Adaptivity of averaged stochastic gradient descent to local strong convexity for logistic regression

Francis Bach

Journal of Machine Learning Research 15 (2014) 367-399 Submitted 10/13; Revised 12/13; Published 2/14

Adaptivity of Averaged Stochastic Gradient Descent to Local Strong Convexity for Logistic Regression

Francis Bach FRANCIS.BACH@ENS.FR
INRIA - Sierra Project-team
Département d'Informatique de l'Ecole Normale Supérieure
Paris, France

Editor: Léon Bottou

Abstract

In this paper, we consider supervised learning problems such as logistic regression and study the stochastic gradient method with averaging, in the usual stochastic approximation setting where observations are used only once. We show that after N iterations, with a constant step-size proportional to $1/R^2\sqrt{N}$ where N is the number of observations and R is the maximum norm of the observations, the convergence rate is always of order $O(1/\sqrt{N})$, and improves to $O(R^2/\mu N)$ where μ is the lowest eigenvalue of the Hessian at the global optimum (when this eigenvalue is greater than $R^2/\sqrt{N}$). Since μ does not need to be known in advance, this shows that averaged stochastic gradient is adaptive to *unknown local* strong convexity of the objective function. Our proof relies on the generalized self-concordance properties of the logistic loss and thus extends to all generalized linear models with uniformly bounded features.

Keywords: stochastic approximation, logistic regression, self-concordance

1. Introduction

The minimization of an objective function which is only available through unbiased estimates of the function values or its gradients is a key methodological problem in many disciplines. Its analysis has been attacked mainly in three scientific communities: stochastic approximation (Fabian, 1968; Ruppert, 1988; Polyak and Juditsky, 1992; Kushner and Yin, 2003; Broadie et al., 2009), optimization (Nesterov and Vial, 2008; Nemirovski et al., 2009), and machine learning (Bottou and Le Cun, 2005; Shalev-Shwartz et al., 2007; Bottou and Bousquet, 2008; Shalev-Shwartz and Srebro, 2008; Shalev-Shwartz et al., 2009; Duchi and Singer, 2009; Xiao, 2010). The main algorithms which have emerged are stochastic gradient descent (a.k.a. Robbins-Monro algorithm), as well as a simple modification where iterates are averaged (a.k.a. Polyak-Ruppert averaging).

For convex optimization problems, the convergence rates of these algorithms depends primarily on the potential *strong convexity* of the objective function (Nemirovski and Yudin, 1983). For μ-strongly convex functions, after n iterations (i.e., n observations), the optimal rate of convergence of function values is $O(1/\mu n)$ while for convex functions the optimal rate is $O(1/\sqrt{n})$, both of them achieved by averaged stochastic gradient with step size respectively proportional to $1/\mu n$ or $1/\sqrt{n}$ (Nemirovski and Yudin, 1983; Agarwal et al.,

2012). For smooth functions, averaged stochastic gradient with step sizes proportional to $1/\sqrt{n}$ achieves them up to logarithmic terms (Bach and Moulines, 2011).

Convex optimization problems coming from supervised machine learning are typically of the form $f(\theta) = \mathbb{E}\big[\ell(y, \langle \theta, x \rangle)\big]$, where $\ell(y, \langle \theta, x \rangle)$ is the loss between the response $y \in \mathbb{R}$ and the prediction $\langle \theta, x \rangle \in \mathbb{R}$, where x is the input data in a Hilbert space $\mathcal{H}$ and linear predictions parameterized by $\theta \in \mathcal{H}$ are considered. They may or may not have strongly convex objective functions. This most often depends on (a) the correlations between covariates x, and (b) the strong convexity of the loss function ℓ. The logistic loss $\ell : u \mapsto \log(1 + e^{-u})$ is not strongly convex unless restricted to a compact set (indeed, restricted to $u \in [-U, U]$, we have $\ell''(u) = e^{-u}(1 + e^{-u})^{-2} \geqslant \frac{1}{4} e^{-U}$). Moreover, in the sequential observation model, the correlations are not known at training time. Therefore, many theoretical results based on strong convexity do not apply (adding a squared norm $\frac{\mu}{2}\|\theta\|^2$ is a possibility, however, in order to avoid adding too much bias, μ has to be small and typically much smaller than $1/\sqrt{n}$, which then makes all strongly-convex bounds vacuous). The goal of this paper is to show that with proper assumptions, namely self-concordance, one can readily obtain favorable theoretical guarantees for logistic regression, namely a rate of the form $O(R^2/\mu n)$ where μ is the lowest eigenvalue of the Hessian at the global optimum, *without any exponentially increasing constant factor* (e.g., with the notations above, without terms of the form e^U).

Another goal of this paper is to design an algorithm and provide an analysis that benefit from *hidden* local strong convexity without requiring to know the local strong convexity constant in advance. In smooth situations, the results of Bach and Moulines (2011) imply that the averaged stochastic gradient method with step sizes of the form $O(1/\sqrt{n})$ is adaptive to the strong convexity of the problem. However the dependence in μ in the strongly convex case is of the form $O(1/\mu^2 n)$, which is sub-optimal. Moreover, the final rate is rather complicated, notably because all possible step-sizes are considered. Finally, it does not apply here because even in low-correlation settings, the objective function of logistic regression cannot be globally strongly convex.

In this paper, we provide an analysis for stochastic gradient with averaging for generalized linear models such as logistic regression, with a step size proportional to $1/R^2\sqrt{n}$ where R is the radius of the data and n the number of observations, showing such adaptivity. In particular, we show that the algorithm can adapt to the *local* strong-convexity constant, that is, the lowest eigenvalue of the Hessian at the optimum. The analysis is done for a finite horizon N and a constant step size decreasing in N as $1/R^2\sqrt{N}$, since the analysis is then slightly easier, though (a) a decaying stepsize could be considered as well, and (b) it could be classically extended to varying step-sizes by a doubling trick (Hazan and Kale, 2001).

2. Stochastic Approximation for Generalized Linear Models

In this section, we present the assumptions our work relies on, as well as related work.

2.1 Assumptions

Throughout this paper, we make the following assumptions. We consider a function f defined on a Hilbert space $\mathcal{H}$, equipped with a norm $\|\cdot\|$. Throughout the paper, we identify the Hilbert space and its dual; thus, the gradients of f also belongs to $\mathcal{H}$ and we

use the same norm on these. Moreover, we consider an increasing family of σ-fields $(\mathcal{F}_n)_{n\geqslant 1}$ and we assume that we are given a deterministic $\theta_0 \in \mathcal{H}$, and a sequence of functions $f_n : \mathcal{H} \to \mathbb{R}$, for $n \geqslant 1$. We make the following assumptions, for a certain $R > 0$:

(A1) Convexity and differentiability of f: f is convex and three-times differentiable.

(A2) Generalized self-concordance of f (Bach, 2010): for all $\theta_1, \theta_2 \in \mathcal{H}$, the function $\varphi : t \mapsto f\big[\theta_1 + t(\theta_2 - \theta_1)\big]$ satisfies: $\forall t \in \mathbb{R}$, $|\varphi'''(t)| \leqslant R\|\theta_1 - \theta_2\|\varphi''(t)$.

(A3) Attained global minimum: f has a global minimum attained at $\theta_* \in \mathcal{H}$.

(A4) Lipschitz-continuity of f_n and f: all gradients of f and f_n are bounded by R, that is, for all $\theta \in \mathcal{H}$,

$$\|f'(\theta)\| \leqslant R \text{ and } \forall n \geqslant 1,\ \|f_n'(\theta)\| \leqslant R \text{ almost surely.}$$

(A5) Adapted measurability: $\forall n \geqslant 1$, f_n is $\mathcal{F}_n$-measurable.

(A6) Unbiased gradients: $\forall n \geqslant 1$, $\mathbb{E}(f_n'(\theta_{n-1})|\mathcal{F}_{n-1}) = f'(\theta_{n-1})$.

(A7) Stochastic gradient recursion: $\forall n \geqslant 1$, $\theta_n = \theta_{n-1} - \gamma_n f_n'(\theta_{n-1})$, where $(\gamma_n)_{n\geqslant 1}$ is a deterministic sequence.

In this paper, we will also consider the averaged iterate $\bar{\theta}_n = \frac{1}{n}\sum_{k=0}^{n-1}\theta_k$, which may be trivially computed on-line through the recursion $\bar{\theta}_n = \frac{1}{n}\theta_{n-1} + \frac{n-1}{n}\bar{\theta}_{n-1}$.

Among the seven assumptions above, the non-standard one is **(A2)**: the notion of self-concordance is an important tool in convex optimization and in particular for the study of Newton's method (Nesterov and Nemirovskii, 1994). It corresponds to having the third derivative bounded by the $\frac{3}{2}$-th power of the second derivative. For machine learning, Bach (2010) has generalized the notion of self-concordance by removing the $\frac{3}{2}$-th power, so that it is applicable to cost functions arising from probabilistic modeling, as shown below. The key consequence of our notion of self-concordance is a relationship shown in Lemma 9 (Section 5) between the norm of a gradient $\|f'(\theta)\|$ and the excess cost function $f(\theta) - f(\theta_*)$, which is the same than for strongly convex functions, but with the local strong convexity constant rather than the global one (which is equal to zero here).

Our set of assumptions corresponds to the following examples (with i.i.d. data, and $\mathcal{F}_n$ equal to the σ-field generated by $x_1, y_1, \dots, x_n, y_n$):

- **Logistic regression**: $f_n(\theta) = \log(1 + \exp(-y_n\langle x_n, \theta\rangle))$, with data x_n uniformly almost surely bounded by R and $y_n \in \{-1, 1\}$. The norm considered here is also the norm of the Hilbert space. Note that this includes other binary classification losses, such as $f_n(\theta) = -y_n\langle x_n, \theta\rangle + \sqrt{1 + \langle x_n, \theta\rangle^2}$.
- **Generalized linear models with uniformly bounded features**: $f_n(\theta) = -\langle\theta, \Phi(x_n, y_n)\rangle + \log\int h(y)\exp\big(\langle\theta, \Phi(x_n, y)\rangle\big)dy$, with $\Phi(x_n, y) \in \mathcal{H}$ almost surely bounded in norm by R, for all observations x_n and all potential responses y in a measurable space. This includes multinomial regression and conditional random fields (Lafferty et al., 2001).
- **Robust regression**: we may use $f_n(\theta) = \varphi(y_n - \langle x_n, \theta\rangle)$, with $\varphi(t) = \log\cosh t = \log\frac{e^t + e^{-t}}{2}$, with a similar boundedness assumption on x_n.

2.2 Running-time Complexity

The stochastic gradient descent recursion $\theta_n = \theta_{n-1} - \gamma_n f_n'(\theta_{n-1})$ operates in full generality in the potentially infinite-dimensional Hilbert space $\mathcal{H}$. There are two practical set-ups where this recursion can be implemented. When $\mathcal{H}$ is finite-dimensional with dimension d, then the complexity of a single iteration is $O(d)$, and thus $O(dn)$ after n iterations. When $\mathcal{H}$ is infinite-dimensional, the recursion can be readily implemented when (a) all functions f_n depend on one-dimensional projections $\langle x_n, \theta \rangle$, that is, are of the form $f_n(\theta) = \varphi_n\big(\langle x_n, \theta \rangle\big)$ for certain random functions φ_n (e.g., $\varphi_n(u) = \ell(y_n, u)$ in machine learning), and (b) all scalar products $K_{ij} = \langle x_i, x_j \rangle$ between x_i and x_j, for $i, j \geqslant 1$, can be computed. This may be done through the classical application of the "kernel trick" (Schölkopf and Smola, 2001; Shawe-Taylor and Cristianini, 2004): if $\theta_0 = 0$, we may represent θ_n as a linear combination of vectors $x_1, \dots, x_n$, that is, $\theta_n = \sum_{i=1}^n \alpha_i x_i$, and the recursion may be written in terms of the weights α_n, through

$$\alpha_n = -\gamma_n x_n \varphi_n'\bigg(\sum_{i=1}^{n-1} \alpha_i K_{ni} \bigg).$$

A key element to notice here is that without regularization, the weights α_i corresponding to previous observations remain constant. The overall complexity of the algorithm is $O(n^2)$ times the cost of evaluating a single kernel function. See Bordes et al. (2005) and Wang et al. (2012) for approaches aiming at reducing the computational load in this setting. Finally, note that in the kernel setting, the function $f(\theta)$ cannot be strongly convex because the covariance operator of x is typically a compact operator, with a sequence of eigenvalues tending to zero (some regularization is then needed).

3. Related Work

In this section, we review related work, first for non-strongly convex problems then for strongly convex problems.

3.1 Non-strongly-convex Functions

When only convexity of the objective function is assumed, several authors (Nesterov and Vial, 2008; Nemirovski et al., 2009; Shalev-Shwartz et al., 2009; Xiao, 2010) have shown that using a step-size proportional to $1/\sqrt{n}$, *together with some form of averaging*, leads to the minimax optimal rate of $O(1/\sqrt{n})$ (Nemirovski and Yudin, 1983; Agarwal et al., 2012). Without averaging, the known convergences rates are suboptimal, that is, averaging is key to obtaining the optimal rate (Bach and Moulines, 2011). Note that the smoothness of the loss does not change the rate, but may help to obtain better constants, with the potential use of acceleration (Lan, 2012). Recent work (Bach and Moulines, 2013) has considered algorithms which improve on the rate $O(1/\sqrt{n})$ for smooth self-concordant losses, such as the square and logistic losses. Their analysis relies on some of the results proved in this paper (in particular the high-order bounds in Section 4).

The compactness of the domain is often used within the algorithm (by using orthogonal projections) and within the analysis (in particular to optimize the step size and obtain high-probability bounds). In this paper, we do not make such compactness assumptions, since in

a machine learning context, the available bound would be loose and hurt practical performance. Note that the analysis of the related dual averaging methods (Nesterov, 2009; Xiao, 2010) has also been carried without compactness assumptions, and previous analyses would also go through in the same set-up for stochastic mirror descent (Nemirovski and Yudin, 1983), at least for bounds in expectation. In the present paper, we derive higher-order bounds and bounds in high-probability where the lack of compactness is harder to deal with.

Another difference between several analyses is the use of decaying step sizes of the form $\gamma_n \propto 1/\sqrt{n}$ vs. the use of a constant step size of the form $\gamma \propto 1/\sqrt{N}$ for a finite known horizon N of iterations. The use of a "doubling trick" as done by Hazan and Kale (2001) for strongly convex optimization, where a constant step size is used for iterations between 2^p and 2^{p+1}, with a constant that is proportional to $1/\sqrt{2^p}$, would allow to obtain an anytime algorithm from a finite horizon one. In order to simplify our analysis, we only consider a finite horizon N and a constant step-size that will be proportional to $1/\sqrt{N}$.

3.2 Strongly-convex Functions

When the function is μ-strongly convex, that is, $\theta \mapsto f(\theta) - \frac{\mu}{2}\|\theta\|^2$ is convex, there are essentially two approaches to obtaining the minimax-optimal rate of $O(1/\mu n)$ (Nemirovski and Yudin, 1983; Agarwal et al., 2012): (a) using a step size proportional to $1/\mu n$ with averaging for non-smooth problems (Nesterov and Vial, 2008; Nemirovski et al., 2009; Xiao, 2010; Shalev-Shwartz et al., 2009; Duchi and Singer, 2009; Lacoste-Julien et al., 2012) or a step size proportional to $1/(R^2 + n\mu)$ also with averaging, for smooth problems, where R^2 is the smoothness constant of the loss of a single observation (Le Roux et al., 2012); (b) for smooth problems, using longer step-sizes proportional to $1/n^\alpha$ for $\alpha \in (1/2, 1)$ *with* averaging (Polyak and Juditsky, 1992; Ruppert, 1988; Bach and Moulines, 2011).

Note that the often advocated step size, that is, of the form C/n where C is larger than $1/\mu$, leads, without averaging to a convergence rate of $O(1/\mu^2 n)$ (Fabian, 1968; Bach and Moulines, 2011), hence with a worse dependence on μ.

The solution (a) requires to have a good estimate of the strong-convexity constant μ, while the second solution (b) does not require to know such estimate and leads to a convergence rate achieving asymptotically the Cramer-Rao lower bound (Polyak and Juditsky, 1992). Thus, this last solution is adaptive to unknown (but positive) amount of strong convexity. However, unless we take the limiting setting $\alpha = 1/2$, it is not adaptive to lack of strong convexity. While the non-asymptotic analysis of Bach and Moulines (2011) already gives a convergence rate in that situation, the bound is rather complicated and also has a suboptimal dependence on μ. Another goal of this paper is to consider a less general result, but more compact and, as already mentioned, a better dependence on the strong convexity constant μ (moreover, as reviewed below, we consider the *local* strong convexity constant, which is much larger).

Finally, note that unless we restrict the support, the objective function for logistic regression cannot be globally strongly convex (since the Hessian tends to zero when $\|\theta\|$ tends to infinity). In this paper we show that stochastic gradient descent with averaging is adaptive to the *local* strong convexity constant, that is, the lowest eigenvalue of the Hessian

of f at the global optimum, without any exponential terms in RD (which would be present if a compact domain of diameter D was imposed and traditional analyses were performed).

3.3 Adaptivity to Unknown Constants

The desirable property of adaptivity to the difficulty of an optimization problem has also been studied in several settings. Gradient descent with constant step size is for example naturally adaptive to the strong convexity of the problem (see, e.g., Nesterov, 2004). In the stochastic context, Juditsky and Nesterov (2010) provide another strategy than averaging with longer step sizes, but for uniform convexity constants.

4. Non-Strongly Convex Analysis

In this section, we study the averaged stochastic gradient method in the non-strongly convex case, that is, without any (global or local) strong convexity assumptions. We first recall existing results in Section 4.1, that bound the expectation of the excess risk leading to a bound in $O(1/\sqrt{N})$. We then show using martingale moment inequalities how all higher-order moments may be bounded in Section 4.2, still with a rate of $O(1/\sqrt{N})$. However, in Section 4.3, we consider the convergence of the squared gradient, with now a rate of $O(1/N)$. This last result is key to obtaining the adaptivity to local strong convexity in Section 5.

4.1 Existing Results

In this section, we review existing results for Lipschitz-continuous non-strongly convex problems (Nesterov and Vial, 2008; Nemirovski et al., 2009; Shalev-Shwartz et al., 2009; Duchi and Singer, 2009; Xiao, 2010). Note that smoothness is not needed here. We consider a constant step size $\gamma_n = \gamma > 0$, for all $n \geqslant 1$, and we denote by $\bar{\theta}_n = \frac{1}{n}\sum_{k=0}^{n-1} \theta_k$ the averaged iterate.

We prove the following proposition, which provides a bound on the expectation of $f(\bar{\theta}_n) - f(\theta_*)$ that decays at rate $O(\gamma + 1/\gamma n)$, hence the usual choice $\gamma \propto 1/\sqrt{n}$:

Lemma 1 *Assume **(A1)** and **(A3-7)**. With constant step size equal to γ, for any $n \geqslant 0$, we have:*

$$\mathbb{E} f\left(\frac{1}{n}\sum_{k=1}^{n} \theta_{k-1}\right) - f(\theta_*) + \frac{1}{2\gamma n}\mathbb{E}\|\theta_n - \theta_*\|^2 \leqslant \frac{1}{2\gamma n}\|\theta_0 - \theta_*\|^2 + \frac{\gamma}{2}R^2.$$

Proof We have the following recursion, obtained from the Lipschitz-continuity of f_n:

$$\begin{aligned} \|\theta_n - \theta_*\|^2 &= \|\theta_{n-1} - \theta_*\|^2 - 2\gamma\langle \theta_{n-1} - \theta_*, f_n'(\theta_{n-1})\rangle + \gamma^2 \|f_n'(\theta_{n-1})\|^2 \\ &\leqslant \|\theta_{n-1} - \theta_*\|^2 - 2\gamma\langle \theta_{n-1} - \theta_*, f'(\theta_{n-1})\rangle + \gamma^2 R^2 + M_n, \end{aligned}$$

with

$$M_n = -2\gamma\langle \theta_{n-1} - \theta_*, f_n'(\theta_{n-1}) - f'(\theta_{n-1})\rangle.$$

We thus get, using the classical result from convexity $f(\theta_{n-1}) - f(\theta_*) \leqslant \langle \theta_{n-1} - \theta_*, f'(\theta_{n-1})\rangle$:

$$2\gamma\big[f(\theta_{n-1}) - f(\theta_*)\big] \leqslant \|\theta_{n-1} - \theta_*\|^2 - \|\theta_n - \theta_*\|^2 + \gamma^2 R^2 + M_n. \tag{1}$$

Summing over integers less than n, this implies:

$$\frac{1}{n}\sum_{k=0}^{n-1} f(\theta_k) - f(\theta_*) + \frac{1}{2\gamma n}\|\theta_n - \theta_*\|^2 \leqslant \frac{1}{2\gamma n}\|\theta_0 - \theta_*\|^2 + \frac{\gamma}{2}R^2 + \frac{1}{2\gamma n}\sum_{k=1}^{n} M_k.$$

We get the desired result by taking expectation in the last inequality, and using the expectation $\mathbb{E}M_k = \mathbb{E}(\mathbb{E}(M_k|\mathcal{F}_{k-1})) = 0$ and $f\big(\frac{1}{n}\sum_{k=0}^{n-1}\theta_k\big) \leqslant \frac{1}{n}\sum_{k=0}^{n-1} f(\theta_k)$. ■

The following corollary considers a specific choice of the step size (note that the bound is only true for the last iterate):

Corollary 2 *Assume **(A1)** and **(A3-7)**. With constant step size equal to $\gamma = \frac{1}{2R^2\sqrt{N}}$, we have:*

$$\begin{aligned}\forall n \in \{1,\dots,N\}, \quad \mathbb{E}\|\theta_n - \theta_*\|^2 &\leqslant \|\theta_0 - \theta_*\|^2 + \frac{1}{4R^2},\\ \mathbb{E}f\left(\frac{1}{N}\sum_{k=1}^{N}\theta_{k-1}\right) - f(\theta_*) &\leqslant \frac{R^2}{\sqrt{N}}\|\theta_0 - \theta_*\|^2 + \frac{1}{4\sqrt{N}}.\end{aligned}$$

Note that if $\|\theta_0 - \theta_*\|^2$ was known, then a better step-size would be $\gamma = \frac{\|\theta_0-\theta_*\|}{R\sqrt{N}}$, leading to a convergence rate proportional to $\frac{R\|\theta_0-\theta_*\|}{\sqrt{N}}$. However, this requires an estimate (or simply an upper-bound) of $\|\theta_0 - \theta_*\|^2$, which is typically not available.

We are going to improve this result in several ways:

- All moments of $\|\theta_n - \theta_*\|^2$ and $f(\bar{\theta}_n) - f(\theta_*)$ will be bounded, leading to a sub-exponential behavior. Note that we do not assume that the iterates are restricted to a predefined bounded set, which is the usual assumption made to derive tail bounds for stochastic approximation (Nesterov and Vial, 2008; Nemirovski et al., 2009; Kakade and Tewari, 2009).
- We are going to show that the squared norm of the gradient at $\bar{\theta}_n = \frac{1}{n}\sum_{k=1}^{n}\theta_{k-1}$ converges at rate $O(1/n)$, even in the non-strongly convex case. This will allow us to derive finer convergence rates in presence of local strong convexity in Section 5.
- The bounds above do not explicitly depend on the dimension of the problem, however, in practice, the quantity $R^2\|\theta_0-\theta_*\|^2$ typically *implicitly* scales linearly in the problem dimension.

4.2 Higher-Order and Tail Bound

In this section, we prove novel higher-order bounds (see the proof in Appendix C), both for any constant step-sizes and then for the specific choice $\gamma = \frac{1}{2R^2\sqrt{N}}$. This will immediately lead to tail bounds.

Proposition 3 *Assume **(A1)** and **(A3-7)**. With constant step size equal to γ, for any $n \geqslant 0$ and integer $p \geqslant 1$, we have:*

$$\mathbb{E}\Big(2\gamma n\big[f(\bar{\theta}_n) - f(\theta^*)\big] + \|\theta_n - \theta_*\|^2\Big)^p \leqslant \big(3\|\theta_0 - \theta_*\|^2 + 20np\gamma^2R^2\big)^p.$$

Corollary 4 *Assume **(A1)** and **(A3-7)**. With constant step size equal to $\gamma = \frac{1}{2R^2\sqrt{N}}$, for any integer $p \geqslant 1$, we have:*

$$\begin{aligned} \forall n \in \{1, \dots, N\}, \quad \mathbb{E}\|\theta_n - \theta_*\|^{2p} &\leqslant \Big[\frac{1}{R^2}\big(3R^2\|\theta_0 - \theta_*\|^2 + 5p\big)\Big]^p, \\ \mathbb{E}\big[f(\bar{\theta}_N) - f(\theta^*)\big]^p &\leqslant \Big[\frac{1}{\sqrt{N}}\big(3R^2\|\theta_0 - \theta_*\|^2 + 5p\big)\Big]^p. \end{aligned}$$

In Appendix C, we first provide two alternative proofs of the same result: (a) our original somewhat tedious proof based on taking powers of the inequality in Equation (1) and using martingale moment inequalities, (b) a shorter proof later derived by Bach and Moulines (2013), that uses Burkholder-Rosenthal-Pinelis inequality (Pinelis, 1994, Theorem 4.1). We also provide in Appendix C a direct proof of the large deviation bound that we now present.

Having a bound on all moments allows immediately to derive large deviation bounds in the same two cases (by applying Lemma 11 from Appendix A):

Proposition 5 *Assume **(A1)** and **(A3-7)**. With constant step size equal to γ, for any $n \geqslant 0$ and $t \geqslant 0$, we have:*

$$\begin{aligned} \mathbb{P}\Big(f(\bar{\theta}_n) - f(\theta_*) \geqslant 30\gamma R^2 t + \frac{3\|\theta_0 - \theta_*\|^2}{\gamma n}\Big) &\leqslant 2\exp(-t), \\ \mathbb{P}\Big(\|\theta_n - \theta_*\|^2 \geqslant 60 n\gamma^2 R^2 t + 6\|\theta_0 - \theta_*\|^2\Big) &\leqslant 2\exp(-t). \end{aligned}$$

Corollary 6 *Assume **(A1)** and **(A3-7)**. With constant step size equal to $\gamma = \frac{1}{2R^2\sqrt{N}}$, for any $t \geqslant 0$ we have:*

$$\begin{aligned} \mathbb{P}\Big(f(\bar{\theta}_N) - f(\theta_*) \geqslant \frac{15t}{\sqrt{N}} + \frac{6R^2\|\theta_0 - \theta_*\|^2}{\sqrt{N}}\Big) &\leqslant 2\exp(-t), \\ \mathbb{P}\Big(\|\theta_N - \theta_*\|^2 \geqslant 15R^{-2}t + 6\|\theta_0 - \theta_*\|^2\Big) &\leqslant 2\exp(-t). \end{aligned}$$

We can make the following observations:

- The results above are obtained by direct application of Proposition 3. In Appendix C, we also provide an alternative direct proof of a slightly weaker result, which was suggested and outlined by Alekh Agarwal (personal communication), and that uses Freedman's inequality for martingales (Freedman, 1975, Theorem 1.6).
- The results above bounding the norm between the last iterate and a global optimum extend to the averaged iterate.
- The iterates θ_n and $\bar{\theta}_n$ do not necessarily converge to θ_* (note that θ_* may not be unique in general anyway).
- Given that $(\mathbb{E}[f(\bar{\theta}_n) - f(\theta_*)]^p)^{1/p}$ is affine in p, we obtain a subexponential behavior, that is, tail bounds similar to an exponential distribution. The same decay was obtained by Nesterov and Vial (2008) and Nemirovski et al. (2009), but with an extra orthogonal projection step that is equivalent in our setting to know a bound on $\|\theta_*\|$, which is in practice not available.

– The constants in the bounds of of Proposition 3 (and thus other results as well) could clearly be improved. In particular, we have, for $p = 1, 2, 3$ (see proof in Appendix E):

$$\begin{aligned}
\mathbb{E}\Big(2\gamma n\big[f(\bar{\theta}_n) - f(\theta^*)\big] + \|\theta_n - \theta_*\|^2\Big) &\leqslant \|\theta_0 - \theta_*\|^2 + n\gamma^2 R^2, \\
\mathbb{E}\Big(2\gamma n\big[f(\bar{\theta}_n) - f(\theta^*)\big] + \|\theta_n - \theta_*\|^2\Big)^2 &\leqslant \big(\|\theta_0 - \theta_*\|^2 + 9n\gamma^2 R^2\big)^2, \\
\mathbb{E}\Big(2\gamma n\big[f(\bar{\theta}_n) - f(\theta^*)\big] + \|\theta_n - \theta_*\|^2\Big)^3 &\leqslant \big(\|\theta_0 - \theta_*\|^2 + 20n\gamma^2 R^2\big)^3.
\end{aligned}$$

4.3 Convergence of Gradients

In this section, we prove higher-order bounds on the convergence of the gradient, with an improved rate $O(1/n)$ for $\|f'(\bar{\theta}_n)\|^2$. In this section, we will need the self-concordance property in Assumption **(A2)**.

Proposition 7 *Assume **(A1-7)**. With constant step size equal to γ, for any $n \geqslant 0$ and integer p, we have:*

$$\left(\mathbb{E}\left\|f'\left(\frac{1}{n}\sum_{k=1}^{n}\theta_{k-1}\right)\right\|^{2p}\right)^{1/2p} \leqslant \frac{R}{\sqrt{n}}\left[8\sqrt{p} + \frac{4p}{\sqrt{n}} + 40R^2\gamma p\sqrt{n} + \frac{3}{\gamma\sqrt{n}}\|\theta_0 - \theta_*\|^2 + \frac{3}{\gamma R\sqrt{n}}\|\theta_0 - \theta_*\|\right].$$

Corollary 8 *Assume **(A1-7)**. With constant step size equal to $\gamma = \frac{1}{2R^2\sqrt{N}}$, for any integer p, we have:*

$$\left(\mathbb{E}\left\|f'\left(\frac{1}{N}\sum_{k=1}^{N}\theta_{k-1}\right)\right\|^{2p}\right)^{1/2p} \leqslant \frac{R}{\sqrt{N}}\left[8\sqrt{p} + \frac{4p}{\sqrt{n}} + 20p + 6R^2\|\theta_0 - \theta_*\|^2 + 6R\|\theta_0 - \theta_*\|\right].$$

We can make the following observations:

– The squared norm of the gradient $\|f'(\bar{\theta}_N)\|^2$ converges at rate $O(1/N)$.
– Given that $(\mathbb{E}\|f'(\bar{\theta}_N)\|^{2p})^{1/2p}$ is affine in p, we obtain a subexponential behavior for $\|f'(\bar{\theta}_N)\|$, that is, tail bounds similar to an exponential distribution.
– The proof of Proposition 7 makes use of the self-concordance assumption (that allows to upperbound deviations of gradients by deviations of function values) together with the proof technique of Polyak and Juditsky (1992).

5. Self-Concordance Analysis for Strongly-Convex Problems

In the previous section, we have shown that $\|f'(\bar{\theta}_N)\|^2$ is of order $O(1/N)$. If the function f was strongly convex with constant $\mu > 0$, this would immediately lead to the bound $f(\bar{\theta}_N) - f(\theta_*) \leqslant \frac{1}{2\mu}\|f'(\bar{\theta}_N)\|^2$, of order $O(1/\mu N)$. However, because of the Lipschitz-continuity of f on the full Hilbert space $\mathcal{H}$, it cannot be strongly convex. In this section, we show how the self-concordance assumption may be used to obtain the exact same behavior, but with μ replaced by the *local* strong convexity constant, which is more likely to be strictly positive.

The required property is summarized in the following proposition about (generalized) self-concordant function (see proof in Appendix B.1):

Lemma 9 *Let f be a convex three-times differentiable function from $\mathcal{H}$ to $\mathbb{R}$, such that for all $\theta_1, \theta_2 \in \mathcal{H}$, the function $\varphi : t \mapsto f\big[\theta_1 + t(\theta_2 - \theta_1)\big]$ satisfies: $\forall t \in \mathbb{R}$, $|\varphi'''(t)| \leqslant R\|\theta_1 - \theta_2\|\varphi''(t)$. Let θ_* be a global minimizer of f and μ the lowest eigenvalue of $f''(\theta_*)$, which is assumed strictly positive.*

$$\text{If } \frac{\|f'(\theta)\|R}{\mu} \leqslant \frac{3}{4} \text{ , then } \|\theta - \theta_*\|^2 \leqslant 4\frac{\|f'(\theta)\|^2}{\mu^2} \text{ and } f(\theta) - f(\theta_*) \leqslant 2\frac{\|f'(\theta)\|^2}{\mu}.$$

We may now use this proposition for the averaged stochastic gradient. For simplicity, we only consider the step-size $\gamma = \frac{1}{2R^2\sqrt{N}}$, and the last iterate (see proof in Appendix F):

Proposition 10 *Assume **(A1-7)**. Assume $\gamma = \frac{1}{2R^2\sqrt{N}}$. Let $\mu > 0$ be the lowest eigenvalue of the Hessian of f at the unique global optimum θ_*. Then:*

$$\begin{aligned}
\mathbb{E}f(\bar{\theta}_N) - f(\theta_*) &\leqslant \frac{R^2}{N\mu}\Big(5R\|\theta_0 - \theta_*\| + 15\Big)^4, \\
\mathbb{E}\big\|\bar{\theta}_N - \theta_*\big\|^2 &\leqslant \frac{R^2}{N\mu^2}\Big(6R\|\theta_0 - \theta_*\| + 21\Big)^4.
\end{aligned}$$

We can make the following observations:

- The proof relies on Lemma 9 and requires a control of the probability that $\frac{\|f'(\bar{\theta}_N)\|R}{\mu} \leqslant \frac{3}{4}$, which is obtained from Proposition 7.
- We conjecture a bound of the form $\left[\frac{R^2}{N\mu}(\square R\|\theta_0 - \theta_*\| + \triangle\sqrt{p})^4\right]^p$ for the p-th order moment of $f(\bar{\theta}_N) - f(\theta_*)$, for some scalar constants $\square$ and $\triangle$.
- The new bound now has the term $R\|\theta_0 - \theta_*\|$ with a fourth power (compared to the bound in Lemma 1, which has a second power), which typically grows with the dimension of the underlying space (or the slowness of the decay of eigenvalues of the covariance operator when $\mathcal{H}$ is infinite-dimensional). It would be interesting to study whether this dependence can be reduced.
- The key elements in the previous proposition are that (a) the constant μ is the *local* convexity constant, and (b) the step-size does not depend on that constant μ, hence the claimed adaptivity.
- The bounds are only better than the non-strongly-convex bounds from Lemma 1, when the Hessian lowest eigenvalue is large enough, that is, $\mu R^2\sqrt{N}$ larger than a fixed constant.
- In the context of logistic regression, even when the covariance matrix of the inputs is invertible, then the only available lower bound on μ is equal to the lowest eigenvalue of the covariance matrix times $\exp(-R\|\theta_*\|)$, which is exponentially small. However, the previous bound is overly pessimistic since it is based on an upper bound on the largest possible value of $\langle x, \theta_*\rangle$. In practice, the actual value of μ is much larger and only a small constant smaller than the lowest eigenvalue of the covariance matrix. In order to assess if this result can be improved, it is interesting to look at the asymptotic result from Polyak and Juditsky (1992) for logistic regression, which leads to a limit rate of $1/n$ times $\operatorname{tr} f''(\theta_*)^{-1}\big(\mathbb{E}f_n'(\theta_*)f_n'(\theta_*)^\top\big)$; note that this rate holds both for the

stochastic approximation algorithm and for the global optimum of the training cost, using standard asymptotic statistics results (Van der Vaart, 1998). When the model is well-specified, that is, the log-odds ratio of the conditional distribution of the label given the input is linear, then $\mathbb{E}f_n'(\theta_*)f_n'(\theta_*)^\top = \mathbb{E}f_n''(\theta_*) = f''(\theta_*)$, and the asymptotic rate is exactly d/n, where d is the dimension of $\mathcal{H}$ (which has to be finite-dimensional for the covariance matrix to be invertible). It would be interesting to see if making the extra assumption of well-specification, we can also get an improved *non-asymptotic* result. When the model is mis-specified however, the quantity $\mathbb{E}f_n'(\theta_*)f_n'(\theta_*)^\top$ may be large even when $f''(\theta_*)$ is small, and the asymptotic regime does not readily lead to an improved bound.

6. Conclusion

In this paper, we have provided a novel analysis of averaged stochastic gradient for logistic regression and related problems. The key aspects of our result are (a) the adaptivity to local strong convexity provided by averaging and (b) the use of self-concordance to obtain a simple bound that does not involve a term which is explicitly exponential in $R\|\theta_0 - \theta_*\|$, which could be obtained by constraining the domain of the iterates.

Our results could be extended in several ways: (a) with a finite and known horizon N, we considered a constant step-size proportional to $1/R^2\sqrt{N}$; it thus seems natural to study the decaying step size $\gamma_n = O(1/R^2\sqrt{n})$, which should, up to logarithmic terms, lead to similar results—and thus likely provide a solution to a a recently posed open problem for online logistic regression (McMahan and Streeter, 2012); (b) an alternative would be to consider a doubling trick where the step-sizes are piecewise constant; also, (c) it may be possible to consider other assumptions, such as exp-concavity (Hazan and Kale, 2001) or uniform convexity (Juditsky and Nesterov, 2010), to derive similar or improved results. Finally, by departing from a plain averaged stochastic gradient recursion, Bach and Moulines (2013) have considered an online Newton algorithm with the same running-time complexity, which leads to a rate of $O(1/n)$ without strong convexity assumptions for logistic regression (though with additional assumptions regarding the distributions of the inputs). It would be interesting to understand if simple assumptions such as the ones made in the present paper are possible while preserving the improved convergence rate.

Acknowledgments

The author was partially supported by the European Research Council (SIERRA Project), and thanks Simon Lacoste-Julien, Eric Moulines and Mark Schmidt for helpful discussions. Morever, Alekh Agarwal suggested and provided a detailed outline of the proof technique based on Freedman's inequality; this was greatly appreciated.

Appendix A. Probability Lemmas

In this appendix, we prove simple lemmas relating bounds on moments to tail bounds, with the traditional use of Markov's inequality. See more general results by Boucheron et al. (2013).

Lemma 11 *Let X be a non-negative random variable such that for some positive constants A and B, and all $p \in \{1, \dots, n\}$,*

$$\mathbb{E}X^p \leqslant (A + Bp)^p.$$

Then, if $t \leqslant \frac{n}{2}$,

$$\mathbb{P}(X \geqslant 3Bt + 2A) \leqslant 2\exp(-t).$$

Proof We have, by Markov's inequality, for any $p \in \{1, \dots, n\}$:

$$\mathbb{P}(X \geqslant 2Bp + 2A) \leqslant \frac{\mathbb{E}X^p}{(2Bp + 2A)^p} \leqslant \frac{(A + Bp)^p}{(2A + 2Bp)^p} = \exp(-\log(2)p).$$

For $u \in [1, n]$, we consider $p = \lfloor u \rfloor$, so that

$$\mathbb{P}(X \geqslant 2Bu + 2A) \leqslant \mathbb{P}(X \geqslant 2Bp + 2A) \leqslant \exp(-\log(2)p) \leqslant 2\exp(-\log(2)u).$$

We take $t = \log(2)u$ and use $2/\log 2 \leqslant 3$. This is thus valid if $t \leqslant \frac{n}{2}$. ■

Lemma 12 *Let X be a non-negative random variable such that for some positive constants A, B and C, and for all $p \in \{1, \dots, n\}$,*

$$\mathbb{E}X^p \leqslant (A\sqrt{p} + Bp + C)^{2p}.$$

Then, if $t \leqslant n$,

$$\mathbb{P}(X \geqslant (2A\sqrt{t} + 2Bt + 2C)^2) \leqslant 4\exp(-t).$$

Proof We have, by Markov's inequality, for any $p \in \{1, \dots, n\}$:

$$\begin{aligned}\mathbb{P}(X \geqslant (2A\sqrt{p} + 2Bp + 2C)^2) &\leqslant \frac{\mathbb{E}X^p}{(2A\sqrt{p} + 2Bp + 2C)^{2p}} \\ &\leqslant \frac{(A\sqrt{p} + Bp + C)^{2p}}{(2A\sqrt{p} + 2Bp + 2C)^{2p}} \leqslant \exp(-\log(4)p).\end{aligned}$$

For $u \in [1, n]$, we consider $p = \lfloor u \rfloor$, so that

$$\begin{aligned}\mathbb{P}(X \geqslant (2A\sqrt{u} + 2Bu + 2C)^2) &\leqslant \mathbb{P}(X \geqslant (2A\sqrt{u} + 2Bu + 2C)^2) \\ &\leqslant \exp(-\log(2)p) \leqslant 4\exp(-\log(4)u).\end{aligned}$$

We take $t = \log(4)u$ and use $\log 4 \geqslant 1$. This is thus valid if $t \leqslant n$. ■

Appendix B. Self-Concordance Properties

In this appendix, we show two lemmas regarding our generalized notion of self-concordance, as well as Lemma 9. For more details, see Bach (2010) and references therein.

The following lemma provide an upper-bound on a one-dimensional self-concordant function at a given point which is based on the gradient at this point and the value and the Hessian at the global minimum. This is key to going in Section 5 from a convergence of gradients to a convergence of function values.

Lemma 13 *Let $\varphi : [0,1] \to \mathbb{R}$ a strictly convex three-times differentiable function such that for some $S > 0$, $\forall t \in [0,1]$, $|\varphi'''(t)| \leqslant S\varphi''(t)$. Assume $\varphi'(0) = 0$, $\varphi''(0) > 0$. Then:*

$$\frac{\varphi'(1)}{\varphi''(0)} S \geqslant 1 - e^{-S} \text{ and } \varphi(1) \leqslant \varphi(0) + \frac{\varphi'(1)^2}{\varphi''(0)}(1+S).$$

Moreover, if $\alpha = \frac{\varphi'(1)S}{\varphi''(0)} < 1$, then $\varphi(1) \leqslant \varphi(0) + \frac{\varphi'(1)^2}{\varphi''(0)} \frac{1}{\alpha} \log \frac{1}{1-\alpha}$. If in addition $\alpha \leqslant \frac{3}{4}$, then $\varphi(1) \leqslant \varphi(0) + 2\frac{\varphi'(1)^2}{\varphi''(0)}$ and $\varphi''(0) \leqslant 2\varphi'(1)$.

Proof By self-concordance, we obtain that the derivative of $u \mapsto \log \varphi''(u)$ is lower-bounded by $-S$. By integrating between 0 and $t \in [0,1]$, we get

$$\log \varphi''(t) - \log \varphi''(0) \geqslant -St \text{ , that is, } \varphi''(t) \geqslant \varphi''(0) e^{-St}, \tag{2}$$

and by integrating between 0 and 1, we obtain (note that we have assumed $\varphi'(0) = 0$):

$$\varphi'(1) \geqslant \varphi''(0) \frac{1 - e^{-S}}{S}. \tag{3}$$

We then get (with a first inequality from convexity of φ, and the last inequality from $e^S \geqslant 1 + S$):

$$\varphi(1) - \varphi(0) \leqslant \varphi'(1) \leqslant \varphi'(1) \frac{\varphi'(1)}{\varphi''(0)} \frac{S}{1 - e^{-S}} = \frac{\varphi'(1)^2}{\varphi''(0)} \left(S + \frac{S}{e^S - 1} \right) \leqslant \frac{\varphi'(1)^2}{\varphi''(0)}(1+S).$$

Equation (3) implies that $\alpha \geqslant 1 - e^{-S}$, which implies, if $\alpha < 1$, $S \leqslant \log \frac{1}{1-\alpha}$. This implies that

$$\varphi(1) - \varphi(0) \leqslant \varphi'(1) \frac{\varphi'(1)}{\varphi''(0)} \frac{S}{1 - e^{-S}} \leqslant \frac{\varphi'(1)^2}{\varphi''(0)} \frac{1}{\alpha} \log \frac{1}{1-\alpha},$$

using the monotonicity of $S \mapsto \frac{S}{1-e^{-S}}$. Finally the last bounds are a consequence of $\frac{S}{\alpha} \leqslant \frac{1}{\alpha} \log \frac{1}{1-\alpha} \leqslant 2$, which is valid for $\alpha \leqslant \frac{3}{4}$.

Note that in Equation (2), we do consider a lower-bound on the Hessian with an exponential factor e^{-St}. The key feature of using self-concordance properties is to get around this exponential factor in the final bound. ■

The following lemma upper-bounds the remainder in the first-order Taylor expansion of the gradient by the remainder in the first-order Taylor expansion of the function. This is important when function values behave well (i.e., converge to the minimal value) while the iterates may not.

Lemma 14 *Let f be a convex three-times differentiable function from $\mathcal{H}$ to $\mathbb{R}$, such that for all $\theta_1, \theta_2 \in \mathcal{H}$, the function $\varphi : t \mapsto f\big[\theta_1 + t(\theta_2 - \theta_1)\big]$ satisfies: $\forall t \in \mathbb{R}$, $|\varphi'''(t)| \leqslant R\|\theta_1 - \theta_2\|\varphi''(t)$. For any $\theta_1, \theta_2 \in H$, we have:*

$$\big\|f'(\theta_1) - f'(\theta_2) - f''(\theta_2)(\theta_2 - \theta_1)\big\| \leqslant R\big[f(\theta_1) - f(\theta_2) - \langle f'(\theta_2), \theta_2 - \theta_1\rangle\big].$$

Proof For a given $z \in \mathcal{H}$ of unit norm, let $\varphi(t) = \big\langle z, f'\big(\theta_2 + t(\theta_1 - \theta_2)\big) - f'(\theta_2) - tf''(\theta_2)(\theta_2 - \theta_1)\big\rangle$ and $\psi(t) = R\big[f(\theta_2 + t(\theta_1 - \theta_2)) - f(\theta_2) - t\langle f'(\theta_2), \theta_2 - \theta_1\rangle\big]$. We have $\varphi(0) = \psi(0) = 0$. Moreover, we have the following derivatives:

$$\begin{aligned}
\varphi'(t) &= \big\langle z, f''\big(\theta_2 + t(\theta_1 - \theta_2)\big) - f''(\theta_2), \theta_1 - \theta_2\big\rangle \\
\varphi''(t) &= f'''\big(\theta_2 + t(\theta_1 - \theta_2)\big)[z, \theta_1 - \theta_2, \theta_1 - \theta_2] \\
&\leqslant R\|z\|_2 f''\big(\theta_2 + t(\theta_1 - \theta_2)\big)[\theta_1 - \theta_2, \theta_1 - \theta_2], \text{ using the Appendix A of Bach (2010),} \\
&= R\big\langle \theta_2 - \theta_1, f''\big(\theta_2 + t(\theta_1 - \theta_2)\big)(\theta_1 - \theta_2)\big\rangle \\
\psi'(t) &= R\big\langle f'\big(\theta_2 + t(\theta_1 - \theta_2)\big) - f'(\theta_2), \theta_1 - \theta_2\big\rangle \\
\psi''(t) &= R\big\langle \theta_2 - \theta_1, f''\big(\theta_2 + t(\theta_1 - \theta_2)\big)(\theta_1 - \theta_2)\big\rangle,
\end{aligned}$$

where $f'''(\theta)$ is the third order tensor of third derivatives. This leads to $\varphi'(0) = \psi'(0) = 0$ and $\varphi''(t) \leqslant \psi''(t)$. We thus have $\varphi(1) \leqslant \psi(1)$ by integrating twice, which leads to the desired result by maximizing with respect to z. ■

B.1 Proof of Lemma 9

We follow the standard proof techniques in self-concordant analysis and define an appropriate function of a single real variable and apply simple lemmas like the ones above.

Define $\varphi : t \mapsto f\big[\theta_* + t(\theta - \theta_*)\big] - f(\theta_*)$. We have

$$\begin{aligned}
\varphi'(t) &= \big\langle f'\big[\theta_* + t(\theta - \theta_*)\big], \theta - \theta_*\big\rangle \\
\varphi''(t) &= \big\langle \theta - \theta_*, f''\big[\theta_* + t(\theta - \theta_*)\big](\theta - \theta_*)\big\rangle \\
\varphi'''(t) &= f'''\big[\theta_* + t(\theta - \theta_*)\big][\theta - \theta_*, \theta - \theta_*, \theta - \theta_*].
\end{aligned}$$

We thus have: $\varphi(0) = \varphi'(0) = 0$, $0 \leqslant \varphi'(1) = \langle f'(\theta), \theta - \theta_*\rangle \leqslant \|f'(\theta)\|\|\theta - \theta_*\|$, $\varphi''(0) = \langle \theta - \theta_*, f''(\theta_*)(\theta - \theta_*)\rangle \geqslant \mu\|\theta - \theta_*\|^2$, and $\varphi(t) \geqslant 0$ for all $t \in [0, 1]$. Moreover, $\varphi'''(t) \leqslant R\|\theta - \theta_*\|\varphi''(t)$ for all $t \in [0, 1]$, that is, Lemma 13 applies with $S = R\|\theta - \theta_*\|$. This leads to the desired result, with $\alpha = \frac{\varphi'(1)S}{\varphi''(0)} \leqslant \frac{\|f'(\theta)\|R}{\mu}$. Note that we also have (using the second inequality in Lemma 13), for all $\theta \in \mathcal{H}$ (and without any assumption on θ):

$$f(\theta) - f(\theta_*) \leqslant \big(1 + R\|\theta - \theta_*\|\big)\frac{\|f'(\theta)\|^2}{\mu}.$$

Appendix C. Proof of Proposition 3

We provide two alternative proofs of the same result: (a) our original somewhat tedious proof in Appendices C.3 and C.4, based on taking powers of the inequality in Equation (1)

and using martingale moment inequalities, (b) a shorter proof in Appendix C.5, later derived by Bach and Moulines (2013), that uses Burkholder-Rosenthal-Pinelis inequality (Pinelis, 1994, Theorem 4.1). Another proof technique was suggested and outlined by Alekh Agarwal (personal communication), that uses Freedman's inequality for martingales (Freedman, 1975, Theorem 1.6); it allows to directly get a tail bound like in Proposition 5. This proof will be presented in Appendix C.6.

Note that the two shorter proofs currently lead to slightly worse constants (or to extra logarithmic factors), that may be improved with more refined derivations.

All proofs start from a similar martingale set-up that we describe in Appendix C.1 and use an almost-sure bound when p gets large (Appendix C.2).

C.1 Bounding Martingales

From the proof of Lemma 1, we have the recursion:

$$2\gamma\big[f(\theta_{n-1}) - f(\theta_*)\big] + \|\theta_n - \theta_*\|^2 \quad \leqslant \quad \|\theta_{n-1} - \theta_*\|^2 + \gamma^2 R^2 + M_n,$$

with

$$M_n = -2\gamma\langle \theta_{n-1} - \theta_*, f_n'(\theta_{n-1}) - f'(\theta_{n-1})\rangle.$$

This leads to, by summing from 1 to n, and using the convexity of f:

$$2\gamma n f\left(\frac{1}{n}\sum_{k=1}^n \theta_{k-1}\right) - 2\gamma n f(\theta^*) + \|\theta_n - \theta_*\|^2 \leqslant A_n,$$

with

$$A_n = \|\theta_0 - \theta_*\|^2 + n\gamma^2 R^2 + \sum_{k=1}^n M_k \geqslant 0.$$

Note that A_n may also be defined recursively as $A_0 = \|\theta_0 - \theta_*\|^2$ and

$$A_n = A_{n-1} + \gamma^2 R^2 + M_n. \tag{4}$$

The random variables (M_n) and (A_n) satisfy the following properties that will proved useful throughout the proof:

(a) Martingale increment: for all $k \geqslant 1$, $\mathbb{E}(M_k|\mathcal{F}_{k-1}) = 0$. This implies that $S_n = \sum_{k=1}^n M_k$ is a martingale.

(b) Boundedness: $|M_k| \leqslant 4\gamma R\|\theta_{k-1} - \theta_*\| \leqslant 4\gamma R A_{k-1}^{1/2}$ almost surely.

C.2 Almost Sure Bound

In this section, we derive an almost sure bound that will be valid for small n. From the stochastic gradient recursion $\theta_n = \theta_{n-1} - \gamma f_n'(\theta_{n-1})$, we get, using Assumption **(A4)** and the triangle inequality:

$$\|\theta_n - \theta_*\| \leqslant \|\theta_{n-1} - \theta_*\| + \gamma\|f_n'(\theta_{n-1})\| \leqslant \|\theta_{n-1} - \theta_*\| + \gamma R \ \text{ almost surely.}$$

This leads to $\|\theta_n - \theta_*\| \leqslant \|\theta_0 - \theta_*\| + n\gamma R$ for all $n \geqslant 0$. This in turn implies that

$$\begin{aligned}
A_n &\leqslant \|\theta_0 - \theta_*\|^2 + n\gamma^2 R^2 + 4\gamma R \sum_{k=1}^{n} \|\theta_{k-1} - \theta_*\| \text{ using } |M_k| \leqslant 4\gamma R \|\theta_{k-1} - \theta_*\|, \\
&\leqslant \|\theta_0 - \theta_*\|^2 + n\gamma^2 R^2 + 4\gamma R \sum_{k=1}^{n} \big[\|\theta_0 - \theta_*\| + (k-1)\gamma R\big] \text{ using the inequality above,} \\
&\leqslant \|\theta_0 - \theta_*\|^2 + n\gamma^2 R^2 + 4\gamma n R \|\theta_0 - \theta_*\| + 2\gamma^2 R^2 n^2 \\
&\qquad\qquad \text{by summing over the first } n-1 \text{ integers,} \\
&\leqslant \|\theta_0 - \theta_*\|^2 + n\gamma^2 R^2 + 2\gamma^2 n^2 R^2 + 2\|\theta_0 - \theta_*\|^2 + 2\gamma^2 R^2 n^2 \text{ using } ab \leqslant \frac{a^2}{2} + \frac{b^2}{2}, \\
&\leqslant 3\|\theta_0 - \theta_*\|^2 + 5n^2\gamma^2 R^2 \text{ almost surely.} \qquad (5)
\end{aligned}$$

This implies that the bound is shown for all $p \geqslant \frac{n}{4}$.

C.3 Derivation of p-th Order Recursion

The first proof works as follows: (a) derive a recursion between the p-th moments and the lower-order moments (this section) and (c) prove the result by induction on p (Appendix C.4). Note that we have to treat separately small values on n in the recursion, for which we use the almost sure bound from Appendix C.2.

Starting from Equation (4), using the binomial expansion formula, we get:

$$\begin{aligned}
A_n^p &\leqslant \Big(A_{n-1} + \gamma^2 R^2 + M_n\Big)^p = \sum_{k=0}^{p} \binom{p}{k} \big(A_{n-1} + \gamma^2 R^2\big)^{p-k} M_n^k \\
&\leqslant \big(A_{n-1} + \gamma^2 R^2\big)^p + p\big(A_{n-1} + \gamma^2 R^2\big)^{p-1} M_n + \sum_{k=2}^{p} \binom{p}{k} \big(A_{n-1} + \gamma^2 R^2\big)^{p-k} \big(4\gamma R A_{n-1}^{1/2}\big)^k.
\end{aligned}$$

This leads to, using $E(M_n|\mathcal{F}_{n-1}) = 0$, upper bounding $\gamma^2 R^2$ by $4\gamma^2 R^2$, and using the binomial expansion formula several times:

$$\begin{aligned}
\mathbb{E}\big[A_n^p \big| \mathcal{F}_{n-1}\big] &\leqslant \big(A_{n-1} + 4\gamma^2 R^2\big)^p + \sum_{k=2}^{p} \binom{p}{k} \big(A_{n-1} + 4\gamma^2 R^2\big)^{p-k} \big(4\gamma R A_{n-1}^{1/2}\big)^k \\
&= \big(A_{n-1} + 4\gamma^2 R^2 + 4\gamma R A_{n-1}^{1/2}\big)^p - 4\gamma R p \big(A_{n-1} + 4\gamma^2 R^2\big)^{p-1} A_{n-1}^{1/2} \\
&\qquad\qquad \text{by isolating the term } k = 1 \text{ in the binomial formula,} \\
&= \big(A_{n-1}^{1/2} + 2\gamma R\big)^{2p} - 4\gamma R p \big(A_{n-1} + 4\gamma^2 R^2\big)^{p-1} A_{n-1}^{1/2} \\
&= \sum_{k=0}^{2p} \binom{2p}{k} A_{n-1}^{k/2} (2\gamma R)^{2p-k} - 4\gamma R p A_{n-1}^{1/2} \sum_{k=0}^{p-1} \binom{p-1}{k} A_{n-1}^k (2\gamma R)^{2(p-1-k)} \\
&= \sum_{k=0}^{2p} A_{n-1}^{k/2} (2\gamma R)^{2p-k} C_k,
\end{aligned}$$

with the constants C_k defined as:

$$\begin{aligned} C_{2q} &= \binom{2p}{2q} \text{ for } q \in \{0,\dots,p\}, \\ C_{2q+1} &= \binom{2p}{2q+1} - 2p\binom{p-1}{q} \text{ for } q \in \{0,\dots,p-1\}. \end{aligned}$$

In particular, $C_0 = 1$, $C_{2p} = 1$, $C_1 = 0$ and $C_{2p-1} = \binom{2p}{2p-1} - 2p\binom{p-1}{p-1} = 0$.

Our goal is now to bounding the values of C_k to obtain Equation (8) below. This will be done by bounding the odd-indexed element by the even-indexed elements.

We have, for $q \in \{1,\dots,p-2\}$,

$$\begin{aligned} C_{2q+1}\frac{2q+1}{2p-2q-1} &\leqslant \binom{2p}{2q+1}\frac{2q+1}{2p-2q-1} \\ &= \frac{(2p)!}{(2q+1)!(2p-2q-1)!}\frac{2q+1}{2p-2q-1} \\ &= \frac{(2p)!}{(2q)!(2p-2q)!}\frac{2p-2q}{2p-2q-1} = \binom{2p}{2q}\frac{2p-2q}{2p-2q-1}. \end{aligned} \tag{6}$$

For the end of the interval above in q, that is, $q = p-2$, we obtain $C_{2q+1}\frac{2q+1}{2p-2q-1} \leqslant C_{2q}\frac{4}{3}$, while for $q \leqslant p-3$, we obtain $C_{2q+1}\frac{2q+1}{2p-2q-1} \leqslant C_{2q}\frac{6}{5}$.

Moreover, for $q \in \{1,\dots,p-2\}$,

$$\begin{aligned} C_{2q+1}\frac{2p-2q-1}{2q+1} &\leqslant \binom{2p}{2q+1}\frac{2p-2q-1}{2q+1} \\ &= \frac{(2p)!}{(2q+1)!(2p-2q-1)!}\frac{2p-2q-1}{2q+1} \\ &= \frac{(2p)!}{(2q+2)!(2p-2q-2)!}\frac{2q+2}{2q+1} = \binom{2p}{2q+2}\frac{2q+2}{2q+1}. \end{aligned} \tag{7}$$

For the end of the interval above in q, that is, $q = 1$, we obtain $C_{2q+1}\frac{2p-2q-1}{2q+1} \leqslant C_{2q+2}\frac{4}{3}$, while for $q \geqslant 2$, we obtain $C_{2q+1}\frac{2p-2q-1}{2q+1} \leqslant C_{2q+2}\frac{6}{5}$.

We have moreover, by using the bound $2\gamma R A_{n-1}^{1/2} \leqslant \frac{\alpha}{2}(2\gamma R)^2 + \frac{1}{2\alpha}A_{n-1}$ for $\alpha = \frac{2q+1}{2p-2q-1}$:

$$\begin{aligned} & C_{2q+1}A_{n-1}^{q+1/2}(2\gamma R)^{2p-2q-1} \\ = \;& C_{2q+1}A_{n-1}^{q}(2\gamma R)^{2p-2q-2}A_{n-1}^{1/2}(2\gamma R) \\ \leqslant \;& C_{2q+1}A_{n-1}^{q}(2\gamma R)^{2p-2q-2}\frac{1}{2}\left[\frac{2q+1}{2p-2q-1}(2\gamma R)^2 + \frac{2p-2q-1}{2q+1}A_{n-1}\right] \\ = \;& \frac{1}{2}C_{2q+1}\frac{2p-2q-1}{2q+1}A_{n-1}^{q+1}(2\gamma R)^{2p-2q-2} + \frac{1}{2}C_{2q+1}\frac{2q+1}{2p-2q-1}A_{n-1}^{q}(2\gamma R)^{2p-2q}. \end{aligned}$$

By combining the previous inequality with Equation (6) and Equation (7), we get that the terms indexed by $2q+1$ are bounded by the terms indexed by $2q+2$ and $2q$. All terms with $q \in \{2,\dots,p-3\}$ are expanded with constants $\frac{3}{5}$, while for $q = 1$ and $q = p-2$, this is

$\frac{2}{3}$. Overall each even term receives a contribution which is less than $\max\{\frac{6}{5}, \frac{3}{5}+\frac{2}{3}, \frac{2}{3}\} = \frac{19}{15}$. This leads to

$$\sum_{q=1}^{p-2} C_{2q+1} A_{n-1}^{q+1/2} (2\gamma R)^{2p-2q-1} \leqslant \frac{19}{15} \sum_{q=0}^{p-1} C_{2q} A_{n-1}^{q} (2\gamma R)^{2p-2q},$$

leading to the recursion that will allow us to derive our result:

$$\mathbb{E}\big[A_n^p \big| \mathcal{F}_{n-1}\big] \leqslant A_{n-1}^p + \frac{34}{15} \sum_{q=0}^{p-1} \binom{2p}{2q} A_{n-1}^q (2\gamma R)^{2p-2q}. \tag{8}$$

C.4 Proof by Induction

We now proceed by induction on p. If we assume that $\mathbb{E}A_k^q \leqslant \big(3\|\theta_0 - \theta_*\|^2 + kq\gamma^2 R^2 B\big)^q$ for all $q < p$, and a certain B (which we will choose to be equal to 20). We first note that if $n \leqslant 4p$, then from Equation (5), we have

$$\begin{aligned} \mathbb{E}A_n^p &\leqslant \big(3\|\theta_0 - \theta_*\|^2 + 5n^2\gamma^2 R^2\ \big)^p \\ &\leqslant \big(3\|\theta_0 - \theta_*\|^2 + 20np\gamma^2 R^2\ \big)^p. \end{aligned}$$

Thus, we only need to consider $n \geqslant 4p$. We then get from Equation (8):

$$\begin{aligned} \mathbb{E}\|\theta_n - \theta_*\|^{2p} &\leqslant \|\theta_0 - \theta_*\|^{2p} + \frac{34}{15} \sum_{k=0}^{n-1} \sum_{q=0}^{p-1} \binom{2p}{2q} \mathbb{E}A_k^q (2\gamma R)^{2p-2q} \\ &\leqslant \|\theta_0 - \theta_*\|^{2p} + \frac{34}{15} \sum_{k=0}^{n-1} \sum_{q=0}^{p-1} \binom{2p}{2q} \big(3\|\theta_0 - \theta_*\|^2 + kq\gamma^2 R^2 B\big)^q (2\gamma R)^{2p-2q}, \end{aligned}$$

using the induction hypothesis. We may now sum with respect to k:

$$\begin{aligned} &\mathbb{E}\|\theta_n - \theta_*\|^{2p} \\ \leqslant\ & \|\theta_0 - \theta_*\|^{2p} + \frac{34}{15} \sum_{q=0}^{p-1} \binom{2p}{2q} (2\gamma R)^{2p-2q} \sum_{k=0}^{n-1} \big(3\|\theta_0 - \theta_*\|^2 + kq\gamma^2 R^2 B\big)^q \\ \leqslant\ & \|\theta_0 - \theta_*\|^{2p} + \frac{34}{15} \sum_{q=0}^{p-1} \binom{2p}{2q} (2\gamma R)^{2p-2q} \sum_{j=0}^{q} 3^j \|\theta_0 - \theta_*\|^{2j} \binom{q}{j} \big(q\gamma^2 R^2 B\big)^{q-j} \frac{n^{q-j+1}}{q-j+1} \\ & \text{using } \sum_{k=0}^{n-1} k^\alpha \leqslant \frac{n^{\alpha+1}}{\alpha+1} \text{ for any } \alpha > 0, \\ =\ & \|\theta_0 - \theta_*\|^{2p} + \frac{34}{15} \sum_{j=0}^{p-1} 3^j \|\theta_0 - \theta_*\|^{2j} (4\gamma^2 R^2 n)^{p-j} \sum_{q=j}^{p-1} \binom{2p}{2q} \binom{q}{j} (\tfrac{qB}{4})^{q-j} \frac{n^{q-p+1}}{q-j+1}, \end{aligned}$$

by changing the order of summations. We now aim to show that it is less than

$$\Big(3\|\theta_0 - \theta_*\|^2 + kp\gamma^2 R^2 B\Big)^p = 3^p \|\theta_0 - \theta_*\|^{2p} + \sum_{j=0}^{p-1} 3^j \|\theta_0 - \theta_*\|^{2j} (\gamma^2 R^2 n)^{p-j} (Bp)^{p-j} \binom{p}{j}.$$

By comparing all terms in $\|\theta_0 - \theta_*\|^{2j}$, this is true as soon as for all $j \in \{0, \dots, p-1\}$,

$$\frac{34}{15} \sum_{q=j}^{p-1} \binom{2p}{2q} \binom{q}{j} (qB/4)^{q-j} \frac{1}{q-j+1} \frac{1}{n^{p-q-1}} \leqslant (Bp/4)^{p-j} \binom{p}{j}$$

$$\Leftrightarrow \frac{34}{15} \sum_{k=0}^{p-1-j} \binom{2p}{2k+2} \binom{p-1-k}{j} \big((p-1-k)B/4\big)^{p-1-k-j} \frac{1}{p-k-j} \frac{1}{n^k} \leqslant (Bp/4)^{p-j} \binom{p}{j},$$

obtained by using the change of variable $k = p - 1 - q$. This is implied by, using $n \geqslant 4p$:

$$\frac{136}{15} \sum_{k=0}^{p-1-j} B^{-1-k} p^{-k-p+j} \binom{2p}{2k+2} \frac{\binom{p-1-k}{j}}{\binom{p}{j}} \big(p-1-k\big)^{p-1-k-j} \frac{1}{p-k-j} \leqslant 1.$$

By expanding the binomial coefficients and simplifying by $p-k-j$, this is equivalent to

$$\frac{136}{15} \sum_{k=0}^{p-1-j} B^{-1-k} p^{-k-p+j} \binom{2p}{2k+2} \frac{(p-1-k)\cdots(p-k-j+1)}{p\cdots(p-j+1)} \big(p-1-k\big)^{p-1-k-j} \leqslant 1.$$

We may now write

$$\begin{aligned} \frac{(p-1-k)\cdots(p-k-j+1)}{p\cdots(p-j+1)} &= \frac{(p-1-k)!}{(p-k-j)!} \frac{(p-j)!}{p!} = \frac{(p-1-k)!}{p!} \frac{(p-j)!}{(p-k-j)!} \\ &= \frac{(p-j)\cdots(p-k-j+1)}{p\cdots(p-k)}, \end{aligned}$$

so that we only need to show that

$$\frac{136}{15} \sum_{k=0}^{p-1-j} B^{-1-k} p^{-k-p+j} \binom{2p}{2k+2} \frac{(p-j)\cdots(p-k-j+1)}{p\cdots(p-k)} \big(p-1-k\big)^{p-1-k-j} \leqslant 1.$$

We have, by bounding all terms then than p by p:

$$\begin{aligned}
& \frac{136}{15}\sum_{k=0}^{p-1-j} A^{-1-k}p^{-k-p+j}\binom{2p}{2k+2}\frac{(p-j)\cdots(p-k-j+1)}{p\cdots(p-k)}(p-1-k)^{p-1-k-j} \\
\leqslant\ & \frac{136}{15}\sum_{k=0}^{p-1-j} A^{-1-k}p^{-k-p+j}\binom{2p}{2k+2}\frac{p^k}{p\cdots(p-k)}p^{p-1-k-j} \\
=\ & \frac{136}{15}\sum_{k=0}^{p-1-j} A^{-1-k}p^{-k-1}\binom{2p}{2k+2}\frac{1}{p\cdots(p-k)} \\
=\ & \frac{136}{15}\sum_{k=0}^{p-1-j} A^{-1-k}\frac{p^{-k-1}}{(2k+2)!}\frac{2p(2p-1)\cdots(2p-2k-1)}{p\cdots(p-k)} \\
=\ & \frac{136}{15}\sum_{k=0}^{p-1-j} A^{-1-k}\frac{p^{-2-1}2^{2k+2}}{(2k+2)!}\frac{p(p-1/2)\cdots(p-k-1/2)}{p\cdots(p-k)} \\
\leqslant\ & \frac{136}{15}\sum_{k=0}^{p-1-j} A^{-1-k}\frac{2^{2k+2}}{(2k+2)!} \\
& \text{by associating all } 2k+2 \text{ terms in ratios which are all less than 1,} \\
\leqslant\ & \frac{136}{15}\sum_{k=0}^{+\infty}\frac{(2/\sqrt{A})^{2k+2}}{(2k+2)!} = \frac{136}{15}\big[\cosh(2/\sqrt{A})-1\big] < 1 \text{ if } A \leqslant 20.
\end{aligned}$$

We thus get the desired result $\mathbb{E}A_n^p \leqslant \big(3\|\theta_0-\theta_*\|^2 + 20np\gamma^2R^2\big)^p$, and the proposition is proved by induction.

C.5 Alternative Proof Using Burkholder-Rosenthal-Pinelis Inequality

In this section, we present (a slightly modified version of) the proof from Bach and Moulines (2013) which is based on Burkholder-Rosenthal-Pinelis inequality (Pinelis, 1994, Theorem 4.1), which we now recall.

C.5.1 BRP Inequality

Throughout the proof, we use the notation for $X \in \mathcal{H}$ a random vector, and p any real number greater than 1, $\|X\|_p = \big(\mathbb{E}\|X\|^p\big)^{1/p}$. We first recall the Burkholder-Rosenthal-Pinelis (BRP) inequality (Pinelis, 1994, Theorem 4.1). Let $p \in \mathbb{R}$, $p \geqslant 2$ and $(\mathcal{F}_n)_{n\geqslant 0}$ be a sequence of increasing σ-fields, and $(X_n)_{n\geqslant 1}$ an adapted sequence of elements of $\mathcal{H}$, such that $\mathbb{E}\big[X_n|\mathcal{F}_{n-1}\big] = 0$, and $\|X_n\|_p$ is finite. Then,

$$\begin{aligned}
\left\| \sup_{k\in\{1,\dots,n\}} \left\| \sum_{j=1}^k X_j \right\| \right\|_p &\leqslant \sqrt{p}\left\| \sum_{k=1}^n \mathbb{E}\big[\|X_k\|^2|\mathcal{F}_{k-1}\big] \right\|_{p/2}^{1/2} + p\left\| \sup_{k\in\{1,\dots,n\}} \|X_k\| \right\|_p \qquad (9)\\
&\leqslant \sqrt{p}\left\| \sum_{k=1}^n \mathbb{E}\big[\|X_k\|^2|\mathcal{F}_{k-1}\big] \right\|_{p/2}^{1/2} + p\left\| \sup_{k\in\{1,\dots,n\}} \|X_k\|^2 \right\|_{p/2}^{1/2}.
\end{aligned}$$

C.5.2 Proof of Proposition 3 (With Slightly Worse Constants)

We use BRP's inequality in Equation (9) to get, for $p \in [2, n/4]$:

$$\begin{aligned}
\Big\| \sup_{k\in\{0,\dots,n\}} A_k \Big\|_p &\leqslant \|\theta_0 - \theta_*\|^2 + n\gamma^2 R^2 + \sqrt{p}\, \Big\| 16\gamma^2 R^2 \sum_{k=1}^n \|\theta_{k-1} - \theta_*\|^2 \Big\|_{p/2}^{1/2} \\
&\qquad + p\, \Big\| \sup_{k\in\{1,\dots,n\}} 4\gamma R \|\theta_{k-1} - \theta_*\| \Big\|_p \\
&\leqslant \|\theta_0 - \theta_*\|^2 + n\gamma^2 R^2 + \sqrt{p}\, 4\gamma R\sqrt{n} \Big\| \sup_{k\in\{0,\dots,n-1\}} A_k \Big\|_{p/2}^{1/2} \\
&\qquad + p\, 4\gamma R \Big\| \sup_{k\in\{0,\dots,n-1\}} A_k^{1/2} \Big\|_p \\
&\leqslant \|\theta_0 - \theta_*\|^2 + n\gamma^2 R^2 + 4\gamma R \Big\| \sup_{k\in\{0,\dots,n-1\}} A_k \Big\|_{p/2}^{1/2} \big(\sqrt{pn} + p\big).
\end{aligned}$$

Thus if $B = \Big\| \sup_{k\in\{0,\dots,n\}} A_k \Big\|_p$, we have (using $p \leqslant n/4$, which implies $\sqrt{pn} + p \leqslant \frac{3}{2}\sqrt{pn}$):

$$B \leqslant \|\theta_0 - \theta_*\|^2 + n\gamma^2 R^2 + 6\gamma R B^{1/2} \sqrt{pn}.$$

By solving this quadratic inequality, we get:

$$\big(B^{1/2} - 3\gamma R\sqrt{pn}\big)^2 \leqslant \|\theta_0 - \theta_*\|^2 + n\gamma^2 R^2 + 9\gamma^2 R^2 pn,$$

which implies

$$\begin{aligned}
B^{1/2} &\leqslant 3\gamma R\sqrt{pn} + \sqrt{\|\theta_0 - \theta_*\|^2 + n\gamma^2 R^2 + 9\gamma^2 R^2 pn} \\
B &\leqslant 2 \times 9\gamma^2 R^2 pn + 2 \times \big(\|\theta_0 - \theta_*\|^2 + n\gamma^2 R^2 + 9\gamma^2 R^2 pn\big) \\
&\leqslant 40\gamma^2 R^2 pn + 2\|\theta_0 - \theta_*\|^2.
\end{aligned}$$

The previous statement is valid for $p \geqslant 2$ and trivial for $p = 1$. From Appendix C.2, we only need to have the result for $p \leqslant \frac{n}{4}$. Thus the bound is slightly worse (but could be clearly improved with more care, for example, by using induction on n).

C.6 Alternative Proof Using Freedman's Inequality

In the previous section, we have used p-th order moment martingale inequalities that relate the norm of a martingale to the norm of its predictable quadratic variation process. Similar results may be obtained for tail bounds through Freedman's inequality (Freedman, 1975, Theorem 1.6). This proof technique was suggested and outlined by Alekh Agarwal (personal communication).

C.6.1 Freedman's Inequality and Extensions

Let (X_n) be a real-valued martingale increment adapted to the increasing sequence of σ-fields $(\mathcal{F}_n)$, that is, such that $\mathbb{E}(X_n|\mathcal{F}_n) = 0$, that is almost surely bounded, that is, $|X_n| \leqslant R$

almost surely. Let $\Sigma_n = \sum_{k=1}^n \mathbb{E}(X_k^2|\mathcal{F}_{k-1})$ the predictable quadratic variation process. Then for any constants t and σ^2,

$$\mathbb{P}\Big(\max_{k\in\{1,\dots,n\}} \sum_{i=1}^k X_i \geqslant t, \Sigma_n \leqslant \sigma^2 \Big) \leqslant 2\exp\Big(\frac{-t^2}{2(\sigma^2 + Rt/3)} \Big).$$

When (X_n) are independent random variables, this recovers Bernstein's inequality. From this bound, one may derive the following bound (Kakade and Tewari, 2009); with probability $1 - 4(\log n)\delta$, we have:

$$\max_{k\in\{1,\dots,n\}} \sum_{i=1}^k X_i \leqslant \max\Big\{ 2\sqrt{\Sigma_n}, 3R\sqrt{\log\tfrac{1}{\delta}} \Big\} \sqrt{\log\tfrac{1}{\delta}} \leqslant 2\sqrt{\Sigma_n}\sqrt{\log\tfrac{1}{\delta}} + 3R\log\tfrac{1}{\delta}. \qquad (10)$$

Note that the result of Kakade and Tewari (2009) considers only $\sum_{i=1}^n X_i$ rather than $\max_{k\in\{1,\dots,n\}} \sum_{i=1}^k X_i$, but that the extension of their proof is straightforward.

C.6.2 Proof of Proposition 5 (With Slightly Worse Constants and Scalings)

We can now apply the inequality in Equation (10) to (M_n). We have $|M_n| \leqslant 4\gamma R\|\theta_{n-1} - \theta_*\| \leqslant 4\gamma R\big(\|\theta_0 - \theta_*\| + n\gamma R\big)$ almost surely. Moreover, $\mathbb{E}(M_n^2|\mathcal{F}_{n-1}) \leqslant 16\gamma^2R^2\|\theta_{n-1} - \theta_*\|^2 \leqslant 16\gamma^2R^2A_{n-1}$.

This leads to with probability greater than $1 - 4(\log n)\delta$,

$$\begin{aligned}
\max_{k\in\{1,\dots,n\}} A_k &\leqslant \|\theta_0 - \theta_*\|^2 + n\gamma^2R^2 + 8\gamma R\sqrt{\sum_{k=1}^{n-1} A_k}\sqrt{\log\tfrac{1}{\delta}} + 12\gamma R\big(\|\theta_0 - \theta_*\| + n\gamma R\big)\log\tfrac{1}{\delta} \\
&\leqslant \|\theta_0 - \theta_*\|^2 + n\gamma^2R^2 + 8\gamma R\sqrt{n}\max_{k\in\{1,\dots,n\}}\sqrt{A_k}\sqrt{\log\tfrac{1}{\delta}} \\
&\qquad + 12\gamma R\big(\|\theta_0 - \theta_*\| + n\gamma R\big)\log\tfrac{1}{\delta}.
\end{aligned}$$

We may now solve the quadratic inequality in $\max_{k\in\{1,\dots,n\}}\sqrt{A_k}$. This leads to

$$\begin{aligned}
&\Big(\max_{k\in\{1,\dots,n\}} \sqrt{A_k} - 4\gamma R\sqrt{n}\sqrt{\log\tfrac{1}{\delta}} \Big)^2 \\
\leqslant\ & \|\theta_0 - \theta_*\|^2 + n\gamma^2R^2 + 12\gamma R\big(\|\theta_0 - \theta_*\| + n\gamma R\big)\log\tfrac{1}{\delta} + 16\gamma^2R^2n\log\tfrac{1}{\delta} \\
=\ & \|\theta_0 - \theta_*\|^2 + n\gamma^2R^2 + \big(12\gamma R\|\theta_0 - \theta_*\| + 28n\gamma^2R^2\big)\log\tfrac{1}{\delta}.
\end{aligned}$$

Then

$$\begin{aligned}
&\max_{k\in\{1,\dots,n\}} \sqrt{A_k} \\
\leqslant\ & 4\gamma R\sqrt{n}\sqrt{\log\frac{1}{\delta}} + \|\theta_0 - \theta_*\| + \sqrt{n}\gamma R + \sqrt{12\gamma R\|\theta_0 - \theta_*\| + 28n\gamma^2R^2}\ \sqrt{\log\frac{1}{\delta}}
\end{aligned}$$

and

$$\begin{aligned}
&\max_{k\in\{1,\dots,n\}} A_k \\
\leqslant\ & 64\gamma^2R^2n\log\frac{1}{\delta} + 4\|\theta_0-\theta_*\|^2 + 4n\gamma^2R^2 + 4\big(12\gamma R\|\theta_0-\theta_*\| + 28n\gamma^2R^2\big)\,\log\frac{1}{\delta} \\
\leqslant\ & 4\|\theta_0-\theta_*\|^2 + 4n\gamma^2R^2 + \Big(64\gamma^2R^2n + 48\gamma R\|\theta_0-\theta_*\| + 112n\gamma^2R^2\Big)\log\frac{1}{\delta} \\
\leqslant\ & 4\|\theta_0-\theta_*\|^2 + 4n\gamma^2R^2 + \Big(176\gamma^2R^2n + 48\gamma R\|\theta_0-\theta_*\|\Big)\log\frac{1}{\delta}.
\end{aligned}$$

We thus recover a tail bound which is very similar to the one obtained in Proposition 5, with the following differences: the additional term $48\gamma R\|\theta_0-\theta_*\|$ is unimportant because $\gamma = O(N^{-1/2})$; however, because the extension of Freedman's inequality is satisfied with probability $1-4(\log n)\delta$, this proof technique loses a logarithmic factor.

Appendix D. Proof of Proposition 7

The proof is organized in two parts: we first show a bound on the averaged gradient $\frac{1}{n}\sum_{k=1}^n f'(\theta_{k-1})$, then relate it to the gradient at the averaged iterate, that is, $f'\Big(\frac{1}{n}\sum_{k=1}^n\theta_{k-1}\Big)$, using self-concordance.

D.1 Bound on $\frac{1}{n}\sum_{k=1}^n f'(\theta_{k-1})$

We have, following Polyak and Juditsky (1992) and Bach and Moulines (2011):

$$f_n'(\theta_{n-1}) = \frac{1}{\gamma}(\theta_{n-1}-\theta_n),$$

which implies, by summing over all integers between 1 and n:

$$\frac{1}{n}\sum_{k=1}^n f'(\theta_{k-1}) \ = \ \frac{1}{n}\sum_{k=1}^n\big[f'(\theta_{k-1}) - f_k'(\theta_{k-1})\big] + \frac{1}{\gamma n}(\theta_0-\theta_*) + \frac{1}{\gamma n}(\theta_*-\theta_n).$$

We denote $X_k = \frac{1}{n}\big[f'(\theta_{k-1}) - f_k'(\theta_{k-1})\big] \in \mathcal{H}$. We have: $\|X_k\| \leqslant \frac{2R}{n}$ almost surely and $\mathbb{E}(X_k|\mathcal{F}_{k-1}) = 0$, with $\big(\sum_{k=1}^n\mathbb{E}(\|X_k\|^2|\mathcal{F}_{k-1})\big)^{1/2} \leqslant \frac{2R}{\sqrt{n}}$. We may thus apply the Burkholder-Rosenthal-Pinelis inequality (Pinelis, 1994, Theorem 4.1), and get:

$$\left[\mathbb{E}\left\|\frac{1}{n}\sum_{k=1}^n\big[f'(\theta_{k-1}) - f_k'(\theta_{k-1})\big]\right\|^{2p}\right]^{1/2p} \ \leqslant \ 2p\frac{2R}{n} + \sqrt{2p}\frac{2R}{n^{1/2}}.$$

This leads to, using Proposition 3 and Minkowski's inequality:

$$\begin{aligned}
&\left[\mathbb{E}\left\|\frac{1}{n}\sum_{k=1}^n f'(\theta_{k-1})\right\|^{2p}\right]^{1/2p} \\
\leqslant\ & \left[\mathbb{E}\left\|\frac{1}{n}\sum_{k=1}^n \big[f'(\theta_{k-1}) - f_k'(\theta_{k-1})\big]\right\|^{2p}\right]^{1/2p} + \frac{1}{\gamma n}\|\theta_0-\theta_*\| + \frac{1}{\gamma n}\big[\mathbb{E}\|\theta_*-\theta_n\|^{2p}\big]^{1/2p} \\
\leqslant\ & 2p\frac{2R}{n} + \sqrt{2p}\frac{2R}{n^{1/2}} + \frac{1}{\gamma n}\|\theta_0-\theta_*\| + \big[\frac{1}{\gamma n}\sqrt{3\|\theta_0-\theta_*\|^2 + 20np\gamma^2R^2}\big] \\
\leqslant\ & 2p\frac{2R}{n} + \sqrt{2p}\frac{2R}{n^{1/2}} + \frac{1}{\gamma n}\|\theta_0-\theta_*\| + \big[\frac{\sqrt{3}}{\gamma n}\|\theta_0-\theta_*\| + \frac{1}{\gamma n}\sqrt{20np}\gamma R\big] \\
\leqslant\ & \frac{4pR}{n} + \sqrt{2p}\frac{2R}{n^{1/2}} + \frac{2}{\gamma n}\|\theta_0-\theta_*\| + \frac{1}{\gamma n}\sqrt{20np}\gamma R \\
\leqslant\ & \frac{4pR}{n} + \sqrt{p}\frac{R}{\sqrt{n}}\big[2\sqrt{2}+\sqrt{20}\big] + \frac{1+\sqrt{3}}{\gamma n}\|\theta_0-\theta_*\| \\
\leqslant\ & \frac{4pR}{n} + 8\sqrt{p}\frac{R}{\sqrt{n}} + \frac{3}{\gamma n}\|\theta_0-\theta_*\|. \qquad (11)
\end{aligned}$$

D.2 Using Self-Concordance

Using the self-concordance property of Lemma 14 several times, we obtain:

$$\begin{aligned}
&\left\|\frac{1}{n}\sum_{k=1}^n f'(\theta_{k-1}) - f'\left(\frac{1}{n}\sum_{k=1}^n \theta_{k-1}\right)\right\| \\
=\ & \left\|\frac{1}{n}\sum_{k=1}^n \big[f'(\theta_{k-1}) - f'(\theta_*) - f''(\theta_*)(\theta_{k-1}-\theta_*)\big]\right. \\
&\qquad \left. - f'\left(\frac{1}{n}\sum_{k=1}^n \theta_{k-1}\right) + f'(\theta_*) + f''(\theta_*)\left(\frac{1}{n}\sum_{k=1}^n \theta_{k-1} - \theta_*\right)\right\| \\
\leqslant\ & \frac{R}{n}\sum_{k=1}^n \big[f(\theta_{k-1}) - f(\theta_*) - \langle f'(\theta_*), \theta_{k-1}-\theta_*\rangle\big] \\
&\qquad + R\left[f\left(\frac{1}{n}\sum_{k=1}^n \theta_{k-1}\right) - f(\theta_*) + \left\langle f'(\theta_*), \frac{1}{n}\sum_{k=1}^n \theta_{k-1} - \theta_*\right\rangle\right] \\
\leqslant\ & 2R\left(\frac{1}{n}\sum_{k=1}^n f(\theta_{k-1}) - f(\theta_*)\right) \text{ using the convexity of } f.
\end{aligned}$$

This leads to, using Proposition 3:

$$\begin{aligned}
&\left(\mathbb{E}\left\|\frac{1}{n}\sum_{k=1}^n f'(\theta_{k-1}) - f'\left(\frac{1}{n}\sum_{k=1}^n \theta_{k-1}\right)\right\|^{2p}\right)^{1/2p} \\
\leqslant\ & 2R\left(\mathbb{E}\left[\frac{1}{n}\sum_{k=1}^n f(\theta_{k-1}) - f(\theta_*)\right]^{2p}\right)^{1/2p} \leqslant \frac{2R}{2\gamma n}\left(3\|\theta_0-\theta_*\|^2 + 40np\gamma^2R^2\right). \qquad (12)
\end{aligned}$$

Summing Equation (11) and Equation (12) leads to the desired result.

Appendix E. Results for Small p

In Proposition 3, we may replace the bound $3\|\theta_0 - \theta_*\|^2 + 20np\gamma^2R^2$ with a bound with smaller constants for $p = 1, 2, 3$ (to be used in proofs of results in Section 5). This is done using the same proof principle but finer derivations, as follows. We denote $\gamma^2R^2 = b$ and $\|\theta - \theta_*\|^2 = a$, and consider the following inequalities which we have considered in the proof of Proposition 3:

$$\begin{aligned} A_n^p &\leqslant (A_{n-1} + b + M_n)^p \\ M_n &\leqslant 4b^{1/2}A_{n-1}^{1/2} \text{ and } \mathbb{E}(M_n|\mathcal{F}_{n-1}) = 0, \\ A_0 &= a. \end{aligned}$$

We simply take expansions of the p-th power above, and sum for all first integers. We have:

$$\begin{aligned} \mathbb{E}A_n &\leqslant \mathbb{E}A_{n-1} + b \leqslant a + nb, \\ \mathbb{E}A_n^2 &\leqslant \mathbb{E}(A_{n-1}^2 + b^2 + 2bA_{n-1} + M_n^2) \leqslant \mathbb{E}A_{n-1}^2 + 2\mathbb{E}A_{n-1}b + b^2 + 16b\mathbb{E}A_{n-1} \\ &\leqslant a^2 + 18b\Big[\sum_{k=0}^{n-1} a + kb\Big] + b^2n \leqslant a^2 + 18b[na + \frac{n^2}{2}b] + b^2n \\ &\quad \text{using the result about } \mathbb{E}A_{n-1}, \\ &= a^2 + 18bna + b^2(n + 9n^2) \\ &\leqslant (a + 9nb)^2. \end{aligned}$$

We may now pursue for the third order moments:

$$\begin{aligned} \mathbb{E}A_n^3 &\leqslant \mathbb{E}(A_{n-1} + b)^3 + 3\mathbb{E}(A_{n-1} + b)^2M_n^2 + 3\mathbb{E}(A_{n-1} + b)^3M_n + \mathbb{E}M_{n-1}^3 \\ &\leqslant \mathbb{E}(A_{n-1} + b)^3 + 3\mathbb{E}(A_{n-1} + b)^216bA_{n-1} + 0 + 64b^{3/2}\mathbb{E}A_{n-1}^{3/2} \\ &\leqslant (\mathbb{E}A_{n-1}^3 + 3\mathbb{E}A_{n-1}^2b + 3\mathbb{E}A_{n-1}b^2 + b^3) + 3\mathbb{E}(A_{n-1} + b)16bA_{n-1} + 64b^{3/2}\mathbb{E}A_{n-1}^{3/2} \\ &= (\mathbb{E}A_{n-1}^3 + 3\mathbb{E}A_{n-1}^2b + 3\mathbb{E}A_{n-1}b^2 + b^3) + 3\mathbb{E}(A_{n-1} + b)16bA_{n-1} \\ &\qquad + 32b\mathbb{E}A_{n-1}[2b^{1/2}A_{n-1}^{1/2}]. \end{aligned}$$

By expanding, we get

$$\begin{aligned} \mathbb{E}A_n^3 &\leqslant (\mathbb{E}A_{n-1}^3 + 3\mathbb{E}A_{n-1}^2b + 3\mathbb{E}A_{n-1}b^2 + b^3) + 3\mathbb{E}(A_{n-1} + b)16bA_{n-1} \\ &\qquad + 32\mathbb{E}bA_{n-1}[\frac{A_{n-1}}{4} + 4b] \\ &= EA_{n-1}^3 + \mathbb{E}A_{n-1}^2b[3 + 48 + 8] + \mathbb{E}A_{n-1}b^2[3 + 48 + 128] + b^3 \\ &= \mathbb{E}A_{n-1}^3 + 59\mathbb{E}A_{n-1}^2b + 179\mathbb{E}A_{n-1}b^2 + b^3 \\ &\leqslant a^3 + 59b\Big[\sum_{k=1}^{n-1} a^2 + 18bka + b^2(k + 9k^2)\Big] + 179b^2\Big[\sum_{k=1}^{n-1} a + kb\Big] + nb^3 \\ &\leqslant a^3 + 59b[na^2 + 9bn^2a + b^2(n^2/2 + 3n^3)] + 179b^2[na + bn^2/2] + nb^3 \\ &= a^3 + 59nba^2 + b^2a[59 \cdot 9n^2 + 179n] + b^3[59/2 \cdot n^2 + 3 \cdot 59n^3 + 179/2 \cdot n^2 + n] \\ &= a^3 + 59nba^2 + b^2a[531n^2 + 179n] + b^3[119n^2 + 177n^3 + n] \\ &\leqslant (a + 20nb)^3. \end{aligned}$$

We then obtain:

$$\begin{aligned}\mathbb{E}\left[2\gamma n\big[f(\bar\theta_n)-f(\theta^*)\big]+\|\theta_n-\theta_*\|^2\right]^2 &\leqslant \big(\|\theta_0-\theta_*\|^2+9n\gamma^2R^2\big)^2\\ \mathbb{E}\left[2\gamma n\big[f(\bar\theta_n)-f(\theta^*)\big]+\|\theta_n-\theta_*\|^2\right]^3 &\leqslant \big(\|\theta_0-\theta_*\|^2+20n\gamma^2R^2\big)^3.\end{aligned}$$

Appendix F. Proof of Proposition 10

The proof follows from applying self-concordance properties (Lemma 9) to $\bar\theta_n$. We thus need to provide a control on the probability that $\|f'(\bar\theta_n)\| \geqslant \frac{3\mu}{4R}$.

F.1 Tail Bound for $\|f'(\bar\theta_n)\|$

We derive a large deviation bound, as a consequence of the bound on all moments of $\|f'(\bar\theta_n)\|$ (Proposition 7) and Lemma 12, that allows to go from moments to tail bounds:

$$\mathbb{P}\left(\big\|f'(\bar\theta_n)\big\| \geqslant \frac{2R}{\sqrt{n}}\left[10\sqrt{t}+40R^2\gamma t\sqrt{n}+\frac{3}{\gamma\sqrt{n}}\|\theta_0-\theta_*\|^2+\frac{3}{\gamma R\sqrt{n}}\|\theta_0-\theta_*\|\right]\right)\leqslant 4\exp(-t).$$

In order to derive the bound above, we need to assume that $p \leqslant n/4$ (so that $4p/n \leqslant 2\sqrt{p}/\sqrt{n}$), and thus, when applying Lemma 12, the bound above is valid as long as $t \leqslant n/4$. It is however valid for all t, because the gradients are bounded by R, and for $t > n$, we have $\frac{2R}{\sqrt{n}}10\sqrt{t} \geqslant R$, and the inequality is satisfied with zero probability.

F.2 Bounding the Function Values

From Lemma 9, if $\|f'(\bar\theta_n)\| \geqslant \frac{3\mu}{4R}$, then $f(\bar\theta_n)-f(\theta_*) \leqslant 2\frac{\|f'(\bar\theta_n)\|^2}{\mu}$. This will allow us to derive a tail bound for $f(\bar\theta_n)-f(\theta_*)$, for sufficiently small deviations. For larger deviations, we will use the tail bound which does not use strong convexity (Proposition 5).

We consider the event

$$A_t = \left\{\big\|f'(\bar\theta_n)\big\| \leqslant \frac{2R}{\sqrt{n}}\left[10\sqrt{t}+40R^2\gamma t\sqrt{n}+\frac{3}{\gamma\sqrt{n}}\|\theta_0-\theta_*\|^2+\frac{3}{\gamma R\sqrt{n}}\|\theta_0-\theta_*\|\right]\right\}.$$

We make the following two assumptions regarding γ and t:

$$\begin{aligned}10\sqrt{t}+40R^2\gamma t\sqrt{n} &\leqslant \frac{2}{3}\frac{3\mu}{4R}\frac{\sqrt{n}}{2R}=\frac{\mu\sqrt{n}}{4R^2}\\ \text{and } \frac{3}{\gamma\sqrt{n}}\|\theta_0-\theta_*\|^2+\frac{3}{\gamma R\sqrt{n}}\|\theta_0-\theta_*\| &\leqslant \frac{1}{3}\frac{3\mu}{4R}\frac{\sqrt{n}}{2R}=\frac{\mu\sqrt{n}}{8R^2},\end{aligned} \tag{13}$$

so that the upper-bound on $\|f'(\bar\theta_n)\|$ in the definition of A_t is less than $\frac{3\mu}{4R}$ (so that we can apply Lemma 9). We thus have:

$$\begin{aligned}A_t &\subset \left\{f(\bar\theta_n)-f(\theta_*)\leqslant\frac{8R^2}{\mu n}\left[10\sqrt{t}+40R^2\gamma t\sqrt{n}+\frac{3}{\gamma\sqrt{n}}\|\theta_0-\theta_*\|^2+\frac{2}{\gamma R\sqrt{n}}\|\theta_0-\theta_*\|\right]^2\right\}\\ &\subset \left\{f(\bar\theta_n)-f(\theta_*)\leqslant\frac{8R^2}{\mu n}\left[10\sqrt{t}+20\square t+\triangle\right]^2\right\},\end{aligned}$$

with $\square = 2\gamma R^2\sqrt{n}$ and $\triangle = \frac{3}{\gamma\sqrt{n}}\|\theta_0-\theta_*\|^2 + \frac{3}{\gamma R\sqrt{n}}\|\theta_0-\theta_*\|$.

This implies that for all $t \geqslant 0$, such that $10\sqrt{t} + 20\square t \leqslant \frac{\mu\sqrt{n}}{4R^2}$, that is, our assumption in Equation (13), we may apply the tail bound from Appendix F.1 to get:

$$\mathbb{P}\left(f(\bar{\theta}_n) - f(\theta_*) \geqslant \frac{8R^2}{\mu n}\left[10\sqrt{t} + 20\square t + \triangle\right]^2\right) \leqslant 4e^{-t}. \tag{14}$$

Moreover, we have for all $v \geqslant 0$ (from Proposition 5):

$$\mathbb{P}\left(f(\bar{\theta}_n) - f(\theta_*) \geqslant 30\gamma R^2 v + \frac{3\|\theta_0-\theta_*\|^2}{\gamma n}\right) \leqslant 2\exp(-v). \tag{15}$$

We may now use the last two inequalities to bound the expectation $\mathbb{E}[f(\bar{\theta}_n) - f(\theta_*)]$.

We first express the expectation as an integral of the tail bound and split it into three parts:

$$\begin{aligned}
\mathbb{E}\big[f(\bar{\theta}_n) - f(\theta_*)\big] &= \int_0^{+\infty} \mathbb{P}\big[f(\bar{\theta}_n) - f(\theta_*) \geqslant u\big]du \\
&= \int_0^{\triangle^2\frac{8R^2}{\mu n}} \mathbb{P}\big[f(\bar{\theta}_n) - f(\theta_*) \geqslant u\big]du \\
&\quad + \int_{\triangle^2\frac{8R^2}{\mu n}}^{\frac{8R^2}{\mu n}\left(\frac{\mu\sqrt{n}}{4R^2}+\triangle\right)^2} \mathbb{P}\big[f(\bar{\theta}_n) - f(\theta_*) \geqslant u\big]du \\
&\quad + \int_{\frac{8R^2}{\mu n}\left(\frac{\mu\sqrt{n}}{4R^2}+\triangle\right)^2}^{+\infty} \mathbb{P}\big[f(\bar{\theta}_n) - f(\theta_*) \geqslant u\big]du.
\end{aligned} \tag{16}$$

We may now bound the three terms separately. For the first integral, we bound the probability by one to get $\displaystyle\int_0^{\triangle^2\frac{8R^2}{\mu n}} \mathbb{P}\big[f(\bar{\theta}_n) - f(\theta_*) \geqslant u\big]du \leqslant \triangle^2\frac{8R^2}{n\mu}$.

For the third term in Equation (16), we use the tail bound in Equation (15) to get

$$\begin{aligned}
&\int_{\frac{8R^2}{\mu n}\left(\frac{\mu\sqrt{n}}{4R^2}+\triangle\right)^2}^{+\infty} \mathbb{P}\big[f(\bar{\theta}_n) - f(\theta_*) \geqslant u\big]du \\
=\ &\int_{\frac{8R^2}{\mu n}\left(\frac{\mu\sqrt{n}}{4R^2}+\triangle\right)^2 - \frac{3}{\gamma n}\|\theta_0-\theta_*\|^2}^{+\infty} \mathbb{P}\left[f(\bar{\theta}_n) - f(\theta_*) \geqslant u + \frac{3}{\gamma n}\|\theta_0-\theta_*\|^2\right]du \\
\leqslant\ &2\int_{\frac{8R^2}{\mu n}\left(\frac{\mu\sqrt{n}}{4R^2}+\triangle\right)^2 - \frac{3}{\gamma n}\|\theta_0-\theta_*\|^2}^{+\infty} \exp\big(-\frac{u}{30\gamma R^2}\big)du.
\end{aligned}$$

We may apply Equation (15) because

$$\frac{8R^2}{\mu n}\big(\frac{\mu\sqrt{n}}{4R^2}+\triangle\big)^2 - \frac{3}{\gamma n}\|\theta_0-\theta_*\|^2 \geqslant \frac{8R^2}{\mu n}\big(\frac{\mu\sqrt{n}}{4R^2}+\triangle\big)^2 - \frac{\mu}{8R^2} \geqslant \frac{8R^2}{\mu n}\big(\frac{\mu\sqrt{n}}{4R^2}\big)^2 - \frac{\mu}{8R^2} = \frac{3\mu}{8R^2} \geqslant 0.$$

We can now compute the bound explicitly to get

$$\begin{aligned}
&\int_{\frac{8R^2}{\mu n}\left(\frac{\mu\sqrt{n}}{4R^2}+\triangle\right)^2}^{+\infty} \mathbb{P}\big[f(\bar\theta_n)-f(\theta_*)\geqslant u\big]du \\
\leqslant\ & 60\gamma R^2\exp\left(\frac{-1}{30\gamma R^2}\left[\frac{8R^2}{\mu n}\big(\frac{\mu\sqrt{n}}{4R^2}+\triangle\big)^2-\frac{3}{\gamma n}\|\theta_0-\theta_*\|^2\right]\right)\leqslant 60\gamma R^2\exp\left(\frac{-1}{30\gamma R^2}\frac{3\mu}{8R^2}\right)\\
\leqslant\ & 60\gamma R^2\exp\left(-\frac{\mu}{80\gamma R^4}\right)\leqslant 60\gamma R^2\frac{80\gamma R^4}{2\mu}\text{ using }e^{-\alpha}\leqslant\frac{1}{2\alpha}\text{ for all }\alpha>0\\
=\ & \frac{2400\gamma^2R^6}{\mu}.
\end{aligned}$$

We now consider the second term in Equation (16) for which we will use Equation (14). We consider the change of variable $u=\frac{8R^2}{\mu n}\Big[10\sqrt{t}+20\square t+\triangle\Big]^2$, for which $u\in\Big[\triangle^2\frac{8R^2}{\mu n},\frac{8R^2}{\mu n}\big(\frac{\mu\sqrt{n}}{4R^2}+\triangle\big)^2\Big]$ implies $t\in[0,+\infty)$. This implies that

$$\begin{aligned}
&\int_{\triangle^2\frac{8R^2}{\mu n}}^{\frac{8R^2}{\mu n}\left(\frac{\mu\sqrt{n}}{4R^2}+\triangle\right)^2}\mathbb{P}\big[f(\bar\theta_n)-f(\theta_*)\geqslant u\big]du\\
\leqslant\ & \int_0^\infty 4e^{-t}d\left(\frac{8R^2}{\mu n}\left[10\sqrt{t}+20\square t+\triangle\right]^2\right)\\
=\ & \frac{32R^2}{\mu n}\int_0^\infty e^{-t}\left(100+400\square^2 2t+400\square\frac{3}{2}t^{1/2}+20\triangle\frac{1}{2}t^{-1/2}+40\triangle\square\right)dt\\
=\ & \frac{32R^2}{\mu n}\left(100\Gamma(1)+400\square^2 2\Gamma(2)+400\square\frac{3}{2}\Gamma(3/2)+20\triangle\frac{1}{2}\Gamma(1/2)+40\triangle\square\Gamma(1)\right)\\
& \text{with }\Gamma\text{ denoting the Gamma function,}\\
=\ & \frac{32R^2}{\mu n}\left(100+400\square^2 2+400\square\frac{3}{2}\frac{1}{2}\sqrt{\pi}+20\triangle\frac{1}{2}\sqrt{\pi}+40\triangle\square\right).
\end{aligned}$$

We may now combine the three bounds to get, from Equation (16),

$$\begin{aligned}
\mathbb{E}\big[f(\bar\theta_n)-f(\theta_*)\big]\ \leqslant\ & \triangle^2\frac{8R^2}{n\mu}+\frac{2400\gamma^2R^6}{\mu}\\
&+\frac{32R^2}{\mu n}\left(100+400\square^2 2+400\square\frac{3}{2}\frac{1}{2}\sqrt{\pi}+20\triangle\frac{1}{2}\sqrt{\pi}+40\triangle\square\right)\\
\leqslant\ & \frac{32R^2}{n\mu}\left[\frac{\triangle^2}{4}+75\gamma^2R^4n+100+800\square^2+300\square\sqrt{\pi}+10\triangle\sqrt{\pi}+40\triangle\square\right].
\end{aligned}$$

For $\gamma=\frac{1}{2R^2\sqrt{N}}$, with $\alpha=R\|\theta_0-\theta_*\|$, $\square=1$ and $\triangle=6\alpha^2+6\alpha$, we obtain

$$\begin{aligned}
\mathbb{E}\big[f(\bar\theta_N)-f(\theta_*)\big]&\leqslant\frac{32R^2}{N\mu}\left[\frac{1}{4}\triangle^2+1451+58\Delta\right]\\
&\leqslant\frac{32R^2}{N\mu}\left[9\alpha^4+18\alpha^3+9\alpha^2+1451+348\alpha^2+348\alpha\right]\\
&\leqslant\frac{R^2}{N\mu}\big(625\alpha^4+7500\alpha^3+33750\alpha^2+67500\alpha+50625\big)=\frac{R^2}{N\mu}\big(5\alpha+15\big)^4.
\end{aligned}$$

Note that the previous bound is only valid if $\frac{3}{\gamma\sqrt{n}}\|\theta_0-\theta_*\|^2+\frac{3}{\gamma R\sqrt{n}}\|\theta_0-\theta_*\|\leqslant\frac{\mu\sqrt{n}}{8R^2}$, that is, under the condition $6R^2\|\theta_0-\theta_*\|^2+6R\|\theta_0-\theta_*\|\leqslant\frac{\mu\sqrt{N}}{8R^2}$. If the condition is not satisfied, then the bound is still valid because of Lemma 1. We thus obtain the desired result.

F.3 Bound on Iterates

Following the same principle as for function values in Appendix F.2, we consider the same event A_t. With the same condition on γ and t, we have:

$$A_t\subset\left\{\|\bar{\theta}_n-\theta_*\|^2\leqslant\frac{16R^2}{\mu^2n}\left[10\sqrt{t}+20\square t+\triangle\right]^2\right\},$$

which leads to the tail bound:

$$\mathbb{P}\left(\|\bar{\theta}_n-\theta_*\|^2\geqslant\frac{16R^2}{\mu^2n}\left[10\sqrt{t}+20\square t+\triangle\right]^2\right)\leqslant 4e^{-t}.$$

We may now split the expectation in three integrals:

$$\begin{aligned}\mathbb{E}\|\bar{\theta}_n-\theta_*\|^2 &= \int_0^{\frac{16R^2}{\mu^2n}\triangle^2}\mathbb{P}\big[\|\bar{\theta}_n-\theta_*\|^2\geqslant u\big]du \\ &\quad+\int_{\frac{16R^2}{\mu^2n}\triangle^2}^{\frac{16R^2}{\mu^2n}\left(\frac{\mu\sqrt{n}}{4R^2}+\triangle\right)^2}\mathbb{P}\big[\|\bar{\theta}_n-\theta_*\|^2\geqslant u\big]du \\ &\quad+\int_{\frac{16R^2}{\mu^2n}\left(\frac{\mu\sqrt{n}}{4R^2}+\triangle\right)^2}^{\infty}\mathbb{P}\big[\|\bar{\theta}_n-\theta_*\|^2\geqslant u\big]du.\end{aligned} \tag{17}$$

The first term in Equation (17) is simply bounded by bounding the tail bound by one (like in the previous section): $\int_0^{\frac{16R^2}{\mu^2n}\triangle^2}\mathbb{P}\big[\|\bar{\theta}_n-\theta_*\|^2\geqslant u\big]du\leqslant\frac{16R^2}{\mu^2n}\triangle^2$. The last integral in Equation (17) may be bounded as follows:

$$\begin{aligned}&\int_{\frac{16R^2}{\mu^2n}\left(\frac{\mu\sqrt{n}}{4R^2}+\triangle\right)^2}^{\infty}\mathbb{P}\big[\|\bar{\theta}_n-\theta_*\|^2\geqslant u\big]du \\ =\;&\mathbb{E}\left[1_{\|\bar{\theta}_n-\theta_*\|^2\geqslant\frac{16R^2}{\mu^2n}\left(\frac{\mu\sqrt{n}}{4R^2}+\triangle\right)^2}\|\bar{\theta}_n-\theta_*\|^2\right] \\ \leqslant\;&\mathbb{P}\left[\|\bar{\theta}_n-\theta_*\|^2\geqslant\frac{16R^2}{\mu^2n}\big(\frac{\mu\sqrt{n}}{4R^2}+\triangle\big)^2\right]^{1/2}\left[\mathbb{E}\big(\|\bar{\theta}_n-\theta_*\|^4\big)\right]^{1/2} \\ &\text{using Cauchy-Schwarz inequality,} \\ \leqslant\;&\mathbb{P}\left[\|\bar{\theta}_n-\theta_*\|^2\geqslant\frac{16R^2}{\mu^2n}\big(\frac{\mu\sqrt{n}}{4R^2}+\triangle\big)^2\right]^{1/2}\left(\|\theta_0-\theta_*\|^2+9\gamma^2nR^2\right)\text{ using Proposition 3.}\end{aligned}$$

Moreover, if we denote by t_0 the largest solution of $10\sqrt{t_0} + 20\square t_0 = \frac{\mu\sqrt{n}}{4R^2}$, we have:

$$\begin{aligned}\sqrt{t_0} &= \frac{-10 + \sqrt{100 + 20\square\frac{\mu\sqrt{n}}{R}}}{40\square} = \frac{-10 + 10\sqrt{1 + 20\square\frac{\mu\sqrt{n}}{100R}}}{40\square} \\ &\geqslant \frac{9}{40\square}\sqrt{20\square\frac{\mu\sqrt{n}}{100R}},\end{aligned}$$

as soon as $20\square\frac{\mu\sqrt{n}}{100R} \geqslant 100$, since if $q \geqslant 100$, $-1 + \sqrt{1+q} \leqslant \frac{9}{10}\sqrt{q}$. This implies that

$$\begin{aligned}&\int_{\frac{16R^2}{\mu^2 n}\left(\frac{\mu\sqrt{n}}{4R^2}+\triangle\right)^2}^{\infty} \mathbb{P}\big[\|\bar{\theta}_n - \theta_*\|^2 \geqslant u\big] du \\ \leqslant\ & \left[4\exp(-t_0)\right]^{1/2}\left(\|\theta_0 - \theta_*\|^2 + 9\gamma^2 nR^2\right) \\ \leqslant\ & \frac{9}{2t_0^2}\left(\|\theta_0 - \theta_*\|^2 + 9\gamma^2 nR^2\right) \text{ using } \exp(-\alpha) \leqslant \frac{9}{16\alpha^2} \text{ for all } \alpha > 0, \\ \leqslant\ & \frac{9}{2}\frac{40^4\square^4 100^2 R^4}{9^4 20^2\square^2\mu^2 n}\left[\frac{9}{4}\square^2/R^2 + \frac{\gamma\sqrt{n}}{3}\triangle\right] \\ \leqslant\ & 686 \times 64\frac{\square^2 R^2}{\mu^2 n}\left[\frac{9}{4}\square^2 + \frac{1}{6}\square\triangle\right].\end{aligned}$$

The second term in Equation (17) is bounded exactly like in Appendix F.2, leading to:

$$\begin{aligned}&\int_{\triangle^2\frac{16R^2}{\mu^2 n}}^{\frac{16R^2}{\mu^2 n}\left(\frac{\mu\sqrt{n}}{4R^2}+\triangle\right)^2} \mathbb{P}\big[\|\bar{\theta}_n - \theta_*\|^2 \geqslant u\big] du \\ \leqslant\ & \int_0^\infty 4e^{-t} d\left(\frac{16R^2}{\mu^2 n}\left[10\sqrt{t} + 20\square t + \triangle\right]^2\right) \\ \leqslant\ & \frac{64R^2}{\mu^2 n}\int_0^\infty e^{-t}\left(100 + 400\square^2 2t + 400\square\frac{3}{2}t^{1/2} + 20\triangle\frac{1}{2}t^{-1/2} + 40\triangle\square\right) dt \\ \leqslant\ & \frac{64R^2}{\mu^2 n}\left(100\Gamma(1) + 400\square^2 2\Gamma(2) + 400\square\frac{3}{2}\Gamma(3/2) + 20\triangle\frac{1}{2}\Gamma(1/2) + 40\triangle\square\Gamma(1)\right) dt \\ \leqslant\ & \frac{64R^2}{\mu^2 n}\left(100 + 400\square^2 2 + 400\square\frac{3}{2}\frac{1}{2}\sqrt{\pi} + 20\triangle\frac{1}{2}\sqrt{\pi} + 40\triangle\square\right).\end{aligned}$$

We can now put all elements together to obtain, from Equation (17):

$$\begin{aligned}&\mathbb{E}\|\bar{\theta}_n - \theta_*\|^2 \\ \leqslant\ & \frac{64R^2}{\mu^2 n}\left(100 + 400\square^2 2 + 400\square\frac{3}{2}\frac{1}{2}\sqrt{\pi} + 20\triangle\frac{1}{2}\sqrt{\pi} + 40\triangle\square\right) \\ &\qquad + \frac{16R^2}{\mu^2 n}\triangle^2 + 686 \times 64\frac{\square^2 R^2}{\mu^2 n}\left[\frac{9}{4}\square^2 + \frac{1}{6}\square\triangle\right] \\ \leqslant\ & \frac{64R^2}{n\mu^2}\left[\frac{1}{4}\triangle^2 + 100 + 800\square^2 + 532\square + 32\triangle + 40\triangle\square + 686\frac{9}{4}\square^4 + 686\frac{\triangle\square^3}{6}\right].\end{aligned}$$

For $\gamma = \frac{1}{2R^2\sqrt{N}}$, with $\alpha = R\|\theta_0 - \theta_*\|$, $\square = 1$ and $\triangle = 6\alpha^2 + 6\alpha$, we get

$$\begin{aligned}
\mathbb{E}\|\bar{\theta}_N - \theta_*\|^2 &\leqslant \frac{8R^2}{N\mu^2}\Big[2\triangle^2 + 8\triangle(32+40+115) + 8(100+800+532+1544)\Big] \\
&\leqslant \frac{8R^2}{N\mu^2}\Big[2\triangle^2 + 1496\triangle + 23808\Big] \\
&\leqslant \frac{8R^2}{N\mu^2}\Big[72\alpha^4 + 144\alpha^3 + 72\alpha^2 + 1496\times 6\alpha^2 + 1496\times 6\alpha + 23808\Big] \\
&\leqslant \frac{R^2}{N\mu^2}\Big[1296\alpha^4 + 18144\alpha^3 + 95256\alpha^2 + 222264\alpha + 194481\Big] = \frac{R^2}{N\mu^2}\big(6\alpha+21\big)^4.
\end{aligned}$$

The previous bound is valid as long as $\frac{\mu\sqrt{N}}{R} \geqslant \frac{10000}{20} = 500$. If it is not satisfied, then Lemma 1 shows that it is still valid.

References

A. Agarwal, P. L. Bartlett, P. Ravikumar, and M. J. Wainwright. Information-theoretic lower bounds on the oracle complexity of stochastic convex optimization. *IEEE Transactions on Information Theory*, 58(5):3235–3249, 2012.

F. Bach. Self-concordant analysis for logistic regression. *Electronic Journal of Statistics*, 4: 384–414, 2010.

F. Bach and E. Moulines. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. In *Advances in Neural Information Processing Systems (NIPS)*, 2011.

F. Bach and E. Moulines. Non-strongly-convex smooth stochastic approximation with convergence rate $O(1/n)$. In *Advances in Neural Information Processing Systems (NIPS)*, 2013.

A. Bordes, S. Ertekin, J. Weston, and L. Bottou. Fast kernel classifiers with online and active learning. *Journal of Machine Learning Research*, 6:1579–1619, 2005.

L. Bottou and O. Bousquet. The tradeoffs of large scale learning. In *Advances in Neural Information Processing Systems (NIPS)*, 2008.

L. Bottou and Y. Le Cun. On-line learning for very large data sets. *Applied Stochastic Models in Business and Industry*, 21(2):137–151, 2005.

S. Boucheron, G. Lugosi, and P. Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 2013.

M. N. Broadie, D. M. Cicek, and A. Zeevi. General bounds and finite-time improvement for stochastic approximation algorithms. Technical report, Columbia University, 2009.

J. Duchi and Y. Singer. Efficient online and batch learning using forward backward splitting. *Journal of Machine Learning Research*, 10:2899–2934, 2009.

V. Fabian. On asymptotic normality in stochastic approximation. *The Annals of Mathematical Statistics*, 39(4):1327–1332, 1968.

D. A. Freedman. On tail probabilities for martingales. *Annals of Probability*, 3(1):100–118, 1975.

E. Hazan and S. Kale. Beyond the regret minimization barrier: an optimal algorithm for stochastic strongly-convex optimization. In *Proceedings of the Conference on Learning Theory (COLT)*, 2001.

A. Juditsky and Y. Nesterov. Primal-dual subgradient methods for minimizing uniformly convex functions. Technical Report 00508933, HAL, 2010.

S. M. Kakade and A. Tewari. On the generalization ability of online strongly convex programming algorithms. In *Advances in Neural Information Processing Systems (NIPS)*, 2009.

H. J. Kushner and G. G. Yin. *Stochastic approximation and recursive algorithms and applications*. Springer-Verlag, second edition, 2003.

S. Lacoste-Julien, M. Schmidt, and F. Bach. A simpler approach to obtaining an $O(1/t)$ convergence rate for projected stochastic subgradient descent. Technical Report 1212.2002, ArXiv, 2012.

J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2001.

G. Lan. An optimal method for stochastic composite optimization. *Mathematical Programming*, 133(1-2):365–397, 2012.

N. Le Roux, M. Schmidt, and F. Bach. A stochastic gradient method with an exponential convergence rate for strongly-convex optimization with finite training sets. In *Advances in Neural Information Processing Systems (NIPS)*, 2012.

H. B. McMahan and M. Streeter. Open problem: Better bounds for online logistic regression. In *COLT/ICML Joint Open Problem Session*, 2012.

A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609, 2009.

A. S. Nemirovski and D. B. Yudin. *Problem complexity and method efficiency in optimization*. Wiley & Sons, 1983.

Y. Nesterov. *Introductory lectures on convex optimization: a basic course*. Kluwer Academic Publishers, 2004.

Y. Nesterov. Primal-dual subgradient methods for convex problems. *Mathematical Programming*, 120(1):221–259, 2009.

Y. Nesterov and A. Nemirovskii. *Interior-point polynomial algorithms in convex programming*. SIAM studies in Applied Mathematics, 1994.

Y. Nesterov and J. P. Vial. Confidence level solutions for stochastic programming. *Automatica*, 44(6):1559–1568, 2008.

I. Pinelis. Optimum bounds for the distributions of martingales in banach spaces. *The Annals of Probability*, 22(4):1679–1706, 1994.

B. T. Polyak and A. B. Juditsky. Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, 30(4):838–855, 1992.

D. Ruppert. Efficient estimations from a slowly convergent Robbins-Monro process. Technical Report 781, Cornell University Operations Research and Industrial Engineering, 1988.

B. Schölkopf and A. J. Smola. *Learning with Kernels*. MIT Press, 2001.

S. Shalev-Shwartz and N. Srebro. SVM optimization: inverse dependence on training set size. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2008.

S. Shalev-Shwartz, Y. Singer, and N. Srebro. Pegasos: Primal estimated sub-gradient solver for svm. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2007.

S. Shalev-Shwartz, O. Shamir, N. Srebro, and K. Sridharan. Stochastic convex optimization. In *Proceedings of the Conference on Learning Theory (COLT)*, 2009.

J. Shawe-Taylor and N. Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2004.

A. W. Van der Vaart. *Asymptotic Statistics*. Cambridge Univ. Press, 1998.

Z. Wang, K. Crammer, and S. Vucetic. Breaking the curse of kernelization: Budgeted stochastic gradient descent for large-scale SVM training. *Journal of Machine Learning Research*, 13:3103–3131, 2012.

L. Xiao. Dual averaging methods for regularized stochastic learning and online optimization. *Journal of Machine Learning Research*, 9:2543–2596, 2010.